DOCUMENT RESUME

ED 186 181 TM 800 202

AUTHOR Secolsky, Charles
TITLE Assessing the Interpretive Component of Criterion-Referenced Test Item Validity.
PUB DATE Apr 80
NOTE 100p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Ambiguity; Criterion Referenced Tests; Electronics; Higher Education; *Item Analysis; Learning Processes; Response Style (Tests); *Student Attitudes; *Testing Problems; *Test Items; *Test Validity
IDENTIFIERS Domain Referenced Tests

ABSTRACT

Undergraduates responded to an objective test in electronics and classified each item by domain (one of 14 topics covered in their text), and by type of knowledge (definition, fact, principle, or interpretation). These judgments were compared to their instructor's "standard" judgments. From these data, an index of item-domain divergence in perceived item meaning between examinees and the content specialist was computed, in addition to traditional item statistics. This new measure provides a means for detecting ambiguous items not otherwise detectable using biserial correlations. Items deemed ambiguous by this procedure were compared with items indicated as being ambiguous by students in taped interviews. Avenues for future research are discussed. (Author/CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED186481

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# ASSESSING THE INTERPRETIVE COMPONENT OF CRITERION-REFERENCED TEST ITEM VALIDITY

by
Charles Secolsky

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Charles Secolsky

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Paper presented at the Annual Meeting of the American Educational Research Association, Boston, April 7 - 11, 1980.

TM 800202

## ABSTRACT

An innovative test technique is proposed for assessing the interpretive component of criterion-referenced test item validity. In addition to responding, undergraduate students in the architecture curriculum classified their test items as having been generated from the list of topics from their text. These judgments were compared to their instructor's "standard" judgments. From these data, an index of item-domain divergence in perceived item meaning between examinees and the content specialist and other exploratory items statistics were computed in addition to traditional item statistics. These new measures provide a means for detecting ambiguous items not otherwise detectable using biserial correlations. Items deemed ambiguous by these procedures were compared with items indicated as being ambiguous by students in taped interviews. Avenues for future research are discussed.

# INTRODUCTION

In the last few years a considerable amount of attention has been given to criterion-referenced testing and measurement with regard to a variety of theoretical and practical problems. The problem of item validity, however, until recently has received only limited attention from educational measurement specialists. This is problematic for the evaluation of learning outcomes using criterion-referenced tests-instruments used to make decisions concerning student mastery of specific instructional objectives. If the items are not measures of the intended instructional objectives, then decision--theoretical approaches for allocating students to mastery status, and sophisticated procedures for determining optimal cut-off scores to name a few important developments--are of questionable value.

The validity of a set of criterion-referenced test items is the determination of what the items measure. It is the interpretation of scores or responses to an item that is validated, not the item per se (American Psychological Association, 1974; Cronbach, 1971; Linn, 1979). The content validity of an item is defined as the degree to which the specification of the domain is clear and the extent to which the item has been generated from the clearly defined domain (Hambleton, 1978; Linn, 1977). When the domain is unclear, the value of criterion-referenced interpretations is limited regardless of the soundness of the interpretation. Hambleton (1978) stated:

> Generally speaking, the quality of criterion-referenced test items can be determined by the extent to which they reflect, in terms of their content, the domains from which they were derived. (p. 13)

In other words, an interpretation based on responses to an item may be valid. The value of this valid interpretation, however, may be limited if the item that elicited the responses lacks content validity. Therefore, content validity is a necessary but not sufficient requirement for item validation. To validate interpretations of criterion-referenced test results it is necessary to proceed beyond the considerations of the content validity of the items (Cronbach, 1971; Hambleton, 1978; Linn, 1979; Messick, 1975). Messick (1975) stated:

> The major problem . . .is that content validity . . . is focused upon test forms rather than test scores, upon instruments rather than measurements. Inferences in educational and psychological measurement are made from scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. Call it 'content relevance,' if you will or 'content representativeness,' but don't call it 'content validity' because it doesn't provide evidence for the interpretation of responses or scores. (p. 960)

A consideration of the construct validity of interpretations of test results is necessary for determining the meaning of the responses.

> More specifically, does the description of the person in terms of the construct together with other information about him in various situations provide implications that are true? (Cronbach, 1971: p. 446)

The extent to which a valid interpretation is valuable, however, rests on the explicitness of the domain definition operationalized by the item.

Therefore, considerations of both content validity and construct validity are important in determining the extent to which a measure is valid. Hambleton (1978) stated:

> A content validity study is essential at the test development stage, and the content validity of a criterion-referenced test will influence the kind of test score interpretations that are possible. Also, it is most important to conduct construct validation studies to validate the intended use of the test scores. The nature of the construct validation studies will depend on the intended use of the test scores. (p. 9)

It is the major contention of this study that there exist a number of potential sources of item invalidity inherent in the responses to the items and not the items themselves which are not detectable nor foreseeable by content specialists who rate items as measures of domain specifications. Content specialists may be blind to certain examinee misinterpretations or misperceptions. Collectively, they may have different ideas about what an item means especially as compared to average and below average students. For only the examinees themselves really know the domain, if any, they each individually judged or perceived an item to have been generated from while in the process of responding to the item. It seems that in order for an interpretation of test results to be more sound or valid, additional information in the form of examinee judgments of the items must be determined independently from their responses to them. Just as content specialists' judgments enhance the knowledge of an item's characteristics (e.g. content relevance, type of knowledge, linguistic complexity), examinee judgments enhance the knowledge about the truth of an

inference made about an examinee with respect to some criterion-referenced attribute. This focus differs from Cronbach's (1971) position that a "judge who knows a great deal about the faults discovered in past tests is likely to detect impurities unsuspected by the developer" (p. 475). It extends Cronbach's position by attending to another significant group of "judges": the examinees.

## The Quality of Criterion-Referenced Test Scores as Indicators of Achievement

The behavior exhibited by an individual responding to the items of a criterion-referenced test is often used to evaluate the performance of an examinee via inferences about achievement. The populations of these inferences are item pools measuring behaviors from well-defined behavior domains. Each item in a pool represents the smallest unit within the domain upon which statistics can be calculated. Most of the statistics for items feature some version of "proportion correct." Inferences based on these statistics seem more appropriate for estimating performance rather than achievement. Furthermore, the expression "well-defined behavior domain" connotes more complexity than just responding correctly or incorrectly to an item. Examinee perceptions of items, for example, have never been included in behavior domain definitions. It seems reasonable then to say that the logical ground upon which one can infer about what another has achieved is shaky--unless, of course, one assumes that the behaviors described in well-defined behavior domains are related to the way knowledge is acquired (see Scandura and Wells, 1967). Therefore, inferences should

be restricted to performance domains as suggested by Linn (1977) and others. However, it is true that until now the knowledge achieved by an examinee could best be described by the well-defined and "pre-ordinate" behavior domains that generated the items which are used to evaluate performance.

The clarity of a criterion-referenced test for determining an examinee's status with respect to the well-defined behavior domain depends on the quality of the domain specifications used to generate the items and describe examinee performance (Popham, 1978). The inferences made about persons with respect to the extent they possess some attribute or characteristic are almost always based upon someone's perception that a certain behavior fits the well-defined domain of behaviors. If an examinee emits incorrect responses to three or more binary, multiple-choice items from the same well-defined behavior domain, then we infer from our observation or perception what he/she does not possess the knowledge to correctly respond to the items which embody that behavior domain. Examinees who incorrectly respond to an item may have acquired the knowledge and emitted correct behaviors in thinking about the question. These behaviors, though, are not specified nor defined and are presently out of the realm of criterion-referenced measurement.

Describing achievement using criterion-referenced tests is only as meaningful as the quality of the definitions of behavior domains. And if behavior domain definitions are insensitive to examinee perceptions of items generated from these definitions, then tests are limited to measuring performance, not achievement. Behavior domain definitions

including examinee perceptions of items for evaluating examinees would better justify inferences about achievement. More would be known about what examinees know. Stake (1967) has advocated the legitimacy of data other than test scores such as the judgments of various constituencies connected with the program under evaluation. House (1973) has referred to this notion as "democratizing knowledge demands" (p. 22).

The way perceptions come about is a complex and idiosyncratic matter. This is the case for an ethnographer's description of some field (see Dawson, 1979) or for examinee judgments of an item's intentions. The evaluator of test results may obtain a false picture of what an individual knows. "The meaning of a work is the integration of its appearance with its content" (House, 1979, p. 9). In the same way, the meaning of a measure is also the integration of its appearance with its content. For criterion-referenced tests, the content of items is provided by the test constructor's specifications of the domain. The appearance is the collection of perceptions of the various audiences concerned with the content of the item. The perceptions of examinees are as important as the perceptions of the test constructor. And when examinees respond to items they do not necessarily reveal all their perceptions. It is not uncommon for an examinee to apply certain knowledge to answer a question and not be able to clearly justify that application. Perceptions of items in responding are almost always restricted to the act of finding the correct answer. However, the meaning that is perceived by examinees includes

some intrinsic description and classification of the images brought to bear on the visually tangible portion of the item. In order for the meaning of criterion-referenced test scores to be more adequately conveyed or understood, one needs more evidence than just objective test scores. An analysis of the images each item evoked in each examinee coupled with their responses to the items would produce a more informative portrayal of the interface of achievement and subject matter. Such analyses would provide test users with the means to see the myriad of the perceptions of an item's intention as a collective in the same way test constructors use the collection of content specialist's ratings to assess item-objective congruence (Rovinelli and Hambleton, 1977). At a basic level this may mean that examinees who know how to divide respond incorrectly to 19/5 while they would have responded correctly to $19 \div 5$. They did not know the symbol "/" as indicating division. At a higher level, it may allow test users to know the likelihood of an examinee who perceived an item about vestigial organs as measuring some aspect of anatomy while it was intended to measure knowledge of some aspect of evolution.

A domain definition is never fully explicit. Analyses of perceptions of items provide, among other things, the basis for determining the likelihood of careless errors and/or partial knowledge, the likelihood of examinee misinterpretation of items, the likelihood of testwiseness, and the basis for understanding the complexity of achievement that must hold if test results are to permit a more coherent interpretation. More

qualitatively sound inferences about items or examinees require some view of the collection of images surrounding examinees' conceptualizations of items. In order to understand or interpret performance on a test to adequately translate the complexity of achievement, more than just the simple proportion of items correct is necessary.

Knowing more about where an item is located in the minds of examinees as compared to the test constructor or content specialists is knowing more about some aspect of an item's validity. If there is considerable divergence about what an item means, then educators should know more about how the item is regarded (e.g., what domain(s) and type(s) of knowledge the item is perceived to be from, different from the domain it was intended). The more the reported similarity between the perceptions of examinees and content specialist about the domain an item is generated from and type of knowledge measured by a criterion-referenced test item, the more sound is the information provided by the test scores alone. However, without knowledge of the nature of the images or perceived intentions of an item on the part of examinees, the soundness of a test score as a representation of achievement cannot be ascertained. The classification of items by examinees would allow test users to know which items are least misconstrued and perhaps more valid. The validation of test items therefore requires an examination of examinee perceptions to help bolster the belief in the test score as an objective measure of achievement.

The most thorough and insightful works about weaknesses in the conception of item validity were done by Lindquist (1933, 1935, 1936). Lindquist and Cook (1933) discussed the shortcomings of the discrimination index as a measure of item validity and described the necessity of a subjective ingredient for an improved measure of item validity that was functionally independent of the discrimination index. This discrimination index was defined as "goodness" and not validity or soundness. If the content of items was not representative of the subject matter tested, and well distributed with reference to the objective of instruction, the index would be less valuable for detecting ambiguous and/or structurally deficient or functionally weak items which lack necessary or contain irrelevant or misleading elements. Unnoticed technical defects in item structure are possibly the principal reasons why many apparently good items fail to function (Lindquist, 1936). Such items may be more often missed by superior than by inferior students. They may be open to more than one interpretation. A response is correct according to the meaning or interpretation intended by the test author (face validity), and incorrect if interpreted according to some unintentional or hidden meaning. In such cases, a greater proportion of lower ability examinees may respond correctly with the intended meaning of the item than the proportion of higher ability examinees even though the higher ability examinees knew the correct response with the intended meaning, but responded to the more obscure interpretation (Lindquist, 1936). Defects are often detected using the discrimination index, but there is no guarantee of this regardless of item difficulty when the index is the biserial r.

Smith and Tyler (1942) described specific techniques for testing knowledge, skill and ability used in the Eight-Year Study. One technique

required students to write out the reasons justifying their written responses to problems. Justifications deemed valid by science teachers were the citing of scientific facts, principles and generalizations. From the other written reasons, evaluators were able to infer certain types of errors embedded in the students' responses. Following this open-ended and subjective form of testing, a more objective form of this technique required students to choose the best solution for each problem and, in addition, select from a list of reasons those necessary to support and thereby validate their response choice. The results of the more objective technique closely paralleled the more open-ended technique. "The advantages of more objective scoring and the possibilities for more extensive sampling of the problem situations led to the adoption of the objective form" (Smith and Tyler, 1942: p. 85). The prespecified topic headings emanating from domain definitions in criterion-referenced testing are analagous to the prespecified list of reasons in the latter study mentioned above. The difference between them is that headings describe universes of content whereas the list of reasons could represent specific applications of principles and are generally more complex.

Until recently, the use of content specialists' judgments has received relatively little attention. Such approaches, however, are now recognized in the field. One method advanced by Rovinelli and Hambleton (1977) for determining the extent to which an item is a measure of an objective is the index of item-objective congruence, $I_{ik}$, which assesses the extent to which several content specialists on the average judge item "k" as representative of the objective "i" it was intended to measure.

($I_{ik}$ ranges from -1 to +1.) This a posteriori procedure assumes that the quality of the items should be determined by their perceived representativeness in a context independent from the process which generated, designed or constructed the items. A second procedure suggested by Hambleton (1978) has content specialists rate the quality of items as measures of intended objectives or domain specifications. Each specialist's ratings are compared to the median responses of the group. A third procedure suggested by Hambleton (1978) is to have content specialists match items to the objectives or domain specifications used to generate the items. Extensions of these procedures can be made to evaluate the relative effectiveness of the content specialists in the rating process.

Polin and Baker (1979) studied test writers' judgment in detail in order to describe human judgment of various qualities or characteristics of test items and to provide an empirical basis to refine the features and details of test specifications. They utilized an eleven point rating scale "as representing a spectrum of probability of match between domain specification and item on the six explicit and two implicit dimensions of item qualities or features" (p. 10). The dimensions are: domain description, content limits, distractor domain or response criterion, format, directions, sample item, linguistic complexity and thinking complexity.

As part of a sociological field research team studying language use and school performance in two school districts in California, MacKay (1974) asked students to read the stimulus words or sentences and state why they had chosen their answers. He then compared the answers and reasons with the test constructors' designation of the correct answers. Responding correctly depended on the student's correct identification of a frame of reference that corresponds to the frame of reference the test

constructor had in mind. Students who chose incorrect answers had demonstrated the abilities the test claimed to measure. MacKay (1974) stated:

> The assumption of one correct answer is based on the test constructor's faith that he and the students share a common symbolic environment in which objects have only one meaning which is apparent to all. In this perspective meaning is given not negotiated and built up over the course of instruction. (p. 233)

According to the authors, these studies attempted to contrast the official version of the meaning of the test items and students' interpretation of the items.

It was assumed in this study that if a student classified an item on both domain and type of knowledge dimensions in the same way it was classified by the content specialist, then the item was more valid than if judgments differed. There was a greater chance that the examinee had not misinterpreted the item. Based on this assumption, the following general sources of item invalidity are advanced:

1. the examinee(s) perceives an item as being generated from a domain different from the intended one due to a misinterpretation of the wording, form or meaning of some aspect of the item.

2. the examinee(s) perceives an item to be measuring either a more or less complex understanding or type of knowledge than intended by the test constructor and/or content specialist(s).

By assessing the degree to which test takers on the average judge an item as having been generated from the domain or type of knowledge it was intended to measure, the soundness of the interpretation of responses (the validity of the measure) is enhanced or lowered through the additional information provided in the form of a judgment measure which is independent from the already possibly contaminated responses themselves. The more similar the perceived intentions of an item between examinees' and content specialists' judgments for item-domain and item-knowledge type dimensions, the more sound and meaningful are the interpretations of the

responses to that item. On the other hand, a lack of perceived similarity between examinees' and content specialists' judgment will make the value of the interpretation of the responses more questionable.

Examinee Misinterpretation of the Domain an Item was Generated From

There are a variety of reasons for an examinee to respond incorrectly to an item. The examinee may have read too deeply into the item or may have misinterpreted the referent of the content represented by the item because of cultural and/or other factors (e.g., not recognizing some symbol necessary for understanding the question). Such reasons for misinterpreting an item seem more likely if the examinee judged the domain an item was generated from different from content specialists. However, it may also be the case that the examinee did not know what the subject matter was about and did not comprehend the item enough to have classified it in the same way as did content specialists.

At least some examinees that respond incorrectly will classify the item in the same way as content specialists. Either an examinee knew the subject matter in a similar enough way to classify the item in the same way, or the examinee guessed "correctly" the domain the item was generated from. Because of the possibility of "correctly" guessing and/or misinterpreting the classification task, the judgment scores for each item will be regarded as containing errors of measurement that are independent of the errors of measurement associated with the responses to that item.

The four possible combinations of response and classification scores for item "k" for examinee "j" ignoring both types of measurement error are shown in Table 1. $I_{ik_1}^{(D)}$, the index of item-domain congruence for content specialists, is defined as the extent to which content specialists on the average judged item "k" as having been generated from domain "i". $I_{ik_2}^{(D)}$, the index of item-domain congruence for examinees, is defined as the extent to which examinees on the average judged item "k" as having been generated from domain "i". $d_{ik}^{(D)}$, the index of item-domain divergence is defined as the difference between indices of item-domain congruence for content specialists and examinees. When there is only one content specialist, $I_{ik_1}^{(D)} = 1$.

TABLE 1

Combinations of Responses to Items and Classifications of Criterion-Referenced Test Item-Domain Congruence with Implications for Item Validity

| Response | Average Domain Judgment Difference, $d_{ik}^{(D)}$, Between Examinee(s) and Content Specialist(s) $d_{ik}^{(D)} = \left\lvert I_{ik_1}^{(D)} - I_{ik_2}^{(D)} \right\rvert$ | Implication for Validity |
|---|---|---|
| Correct | small | valid - correct for right reason |
| Correct | large | possibly less valid - testwiseness |
| Incorrect | small | valid - careless error or partial knowledge |
| Incorrect | large | possibly less valid - misinterpretation |

Increasing the Value of Valid Inferences Using the Interpretive Component of Item Validity

In this case, an objectivism of judgment differences may be possible and useful for confirming the subjective judgments of experts who are experienced with past tests. The important issue, the meaning of a stimulus (an item), has two directly concerned audiences, test makers or users and test takers.

Examinee misclassification of item content was expressed as the average domain judgment difference, $d_{ik}^{(D)}$, between examinees and content specialists, while examinee classifications of knowledge type can be expressed as the average knowledge type judgment difference, $d_{ik}^{(K)}$, between examinees and content specialists, or the difference in the judgments for two kinds of knowledge type divergence measures. $d_{ik}^{(K1)}$ is defined as the average knowledge type judgment difference between examinees and content specialist(s) where the knowledge types are treated only categorically and not hierarchically. $d_{ik}^{(K2)}$ is defined as the average knowledge type judgment difference where the knowledge types are treated hierarchically.

In order to make these average judgment differences more meaningful an examination of the three types of classification, domain, knowledge-type, and hierarchical knowledge-type must be viewed in relation to examinee response data. In this study, this was done only for item-domain average judgment differences for each item.

If an examinee answers an item correctly and classifies it in the same way as content specialists for both item-domain and item-knowledge type dimensions, there is no evidence of misinterpretation in response to

the item. The same is true for an examinee responding incorrectly and classifying the item in the same way as content specialist(s). In the former case, similar item classification is likely needed for responding without misinterpretation. In the latter case, similar item classification is likely related to responding with careless errors or partial or incomplete knowledge.

If, on the other hand, an examinee responds correctly and classifies an item differently from content specialist(s), he/she has likely misinterpreted the classification task, guessed the correct response, or is more test-wise than knowledgeable for that item. However, there is no evidence of misinterpretation in response to the item. Misinterpretation is more likely for both item-domain and item-knowledge type dimensions when an examinee has responded incorrectly and classified differently. However, it is also likely that the examinee may not have classified the item in the same way because of her/his lack of knowledge related to responding incorrectly. The above information is summarized in Table 2.

TABLE 2

Combinations of Responses and Classifications of A Criterion-Referenced Test Item and Likely Reasons for Cell Entries

| | | | |
|---|---|---|---|
| | Incorrect = 0 | Correct = 1 | |
| Same = 1 | Likelihood of careless errors, partial-incomplete knowledge | Likelihood of responding correctly for the right reasons | Number of examinees classifying the same way |
| | Cell-A | Cell-B | (A + B) |
| Y (Classification) | | | |
| Different = 0 | Likelihood of lack of knowledge for classifying the same way, or misinterpretation | Likelihood of misinterpretation of classification task, testwiseness or guessing the correct response | Number of examinees classifying the item a different way |
| | Cell-C | Cell-D | (C + D) |
| | (A + C) | (B + D) | |
| | Number of examinees responding incorrectly | Number of examinees responding correctly | Fixed Total |

## METHODOLOGY

This study was conducted at the end of an instructional unit on electricity in a required undergraduate environmental technology course in the School of Architecture at the University of Illinois at Urbana-Champaign during the Spring, 1979 semester. In addition to answering questions, students were asked to follow the instructions at the top of their test paper. The instructions read, "On the line provided to the right of each item, place the letter(s) of the topic(s) that you think generated that item. The list of topics is projected on the overhead screen." Eighty-one out of eighty-nine examinees at least partially cooperated either at the time of testing or on the following day. In all the test consisted of twenty-one questions with a maximum possible total test score of fifty-four points. The actual test that was administered is included in Appendix A. The topic headings were identical to the section headings of the electricity unit of the course manual written by the professor in charge of the course. Although two teaching assistants wrote the test questions, the professor served as the content specialist by indicating the letter(s) of the topic(s) that generated each of the twenty-one test questions. (The professor's judgments are shown on the line to the right of each question in the actual test in Appendix A.) The list of topic headings as they appeared on the overhead screen appears in Appendix B. The judgments served as the content

specialists' "standard" judgment of domain(s) that generated each item. Thirteen of the twenty-one questions were judged as having been generated from only one content domain or topic heading, six questions were judged as having been generated from two content domains or topics; two questions were judged as having been generated from three domains or topic headings. After the exam was over, the teaching assistants were asked to complete and return a type of knowledge by question table consisting of four types of knowledge and twenty-one questions which contained shortened versions of Easley et al.'s (1967) descriptions of knowledge forms that were previously quoted. The instructions read as follows:

> For each question on the test, place a "+" next to the type(s) of knowledge that describes what the question is asking and use the following descriptions for the types of knowledge:
>
> A. Definition - the key knowledge required is the meaning of a word or words not commonly used;
> B. Fact - knowledge of a particular piece of information that has limited applicability;
> C. Principle - knowledge of broad significance;
> D. Interpretation - ability to understand, interpret, or draw direct conclusions beyond the level of ordinary reading comprehension.

Only one of the two teaching assistants later returned their judgments. These judgments are reported in Appendix C.

On the day following the regular examination period in which the examinees classified the items as having been generated from some domain or set of domains in addition to answering the questions, the students were scheduled to attend four separate laboratory sections with each of

the two teaching assistants as instructors for two sections each. At these laboratory sections, students voluntarily signed up to be interviewed about the test. Of the eighteen students who signed up for interviews, thirteen of them were actually interviewed. (The methodological aspects of the interviewing process are discussed later in this chapter.) Then students were asked to judge the domains from which the questions were generated in case they did not do so during the testing period or if they had not finished classifying all the questions. For those students who had judged items twice and differently during the test and on the following day, the first judgment (made during the test) was retained and the judgment made on the following day was discarded. It was decided that the judgment information obtained during the test period was more relevant to this research since examinees were likely to be more familiar with the way they thought about each question at the time of responding. Therefore, their judgments were more likely to be associated with their answers.

During the same laboratory session, examinees who were present were asked to complete an item information form (see Appendix D). They responded to two questions and then classified the test questions as to the type of knowledge each question was asking in the same way as had the teaching assistant. One question asked the students to describe the manner in which they classified items as to the domain(s) they have been generated from (e.g. answer each question and then classify it before going on to the next

question, answer all the questions first and then went back to classify all the questions, or some combination). Three examinees, one of whom was interviewed, reported that they had classified each question first, and then went on to answer it before proceeding to the next question. The second question asked examinees to describe what they felt was meant by indicating the topics that generated each question. An attempt was made to assess their definition of an item being generated from a topic (e.g. topic(s) that best related to each question, questions that came from a topic or set of topics according to the instruction manual). The bottom portion of the item information form was described as space provided for any comments that may be pertinent to the analysis of the test. Classification of at least some questions according to knowledge type were obtained from forty-four examinees. Forty-three examinees provided classifications according to the type of knowledge for almost all the test questions.

Measures for Determining the Likelihood of Examinee Partial Knowledge, Valid Response, Misinterpretation and Testwiseness

The value of the information provided by examinee judgments for each item is bounded by the extent to which classification scores (1 = same, 0 = different) provide exclusive information not contained in the responses to the items (1 = correct, 0 = incorrect). In order to determine this bound for a particular item, it was necessary to calculate the correlation

between classification scores (Y) and response scores (X). This was done for the two dichotomous variables using $\phi/\phi_{max}$ as described by Glass and Stanley (1970: pp. 161-162) except when the numerators of $\phi$ and $\phi_{max}$ were identical thereby producing a value of 1.0. In such cases, the $\phi$ coefficient was computed between the dichotomous variables Y and X. For a description of $\phi$, the reader is referred to Carroll (1961) and Glass and Stanley (1970: pp. 166-167).

Assuming that the correlation between classification and response is relatively low, the problem of determining the likelihood of misinterpretation remains. For examinees who responded incorrectly and classified differently, one does not know if these individuals lacked the knowledge to comprehend what the item was generally about (e.g. they could not even read the item) or if they genuinely misinterpreted the item. To what extent did examinees in this category really have the knowledge required to respond correctly and misinterpret an invalid item? Or, to what extent were these examinees truly not knowledgeable enough to even understand what the item was about or what types of information or knowledge were necessary to answer the item correctly? In order to determine the likelihood of examinee misinterpretation due to item invalidity, the total test scores of the examinees in Cell C of Table 2 (responded incorrectly and classified differently) was viewed in relation to the distribution of total test scores for all those who responded and classified the item. If the examinees in Cell C for a particular item are the examinees with the lowest total test scores, it is more likely the case that those cell C examinees were not knowledgeable enough to classify the item in the same way as the content specialist. If,

on the other hand, these examinees had relatively high total test scores, it is more likely the case that the examinees misinterpreted the meaning of the item, but had the knowledge required to answer the item correctly had they understood what the item meant as intended by the item writer. A large variance of total test scores in this cell would also imply a greater likelihood of misinterpretation since the responses to the item would not provide good item discrimination with respect to the ability measured by the test as indicated by the biserial r. For this reason, for each item, biserial correlation coefficients were computed separately for those judging the item in the same way as the content specialist and for those judging the item differently from the content specialist.

Again assuming that the absolute value of the correlation between responses and classifications is relatively low, the problem of determining the likelihood of careless errors or partial knowledge and testwiseness in the responses to each item remains. To determine these likelihoods, total test scores are the criterion.

As $\overline{X}_{tot}$ increases for examinees who responded incorrectly to an item and classified it in the same way as the professor, the likelihood that an incorrect response was due to a careless error and that the respondent had partial knowledge increases. However, as $\overline{X}_{tot}$ gets smaller, the likelihood that the item was classified in the "standard way" by guessing increases. Similarly, as $\overline{X}_{tot}$ increases for examinees who responded correctly and classified differently, the likelihood that a correct response was due to testwiseness increases. On the other hand, as $\overline{X}_{tot}$ decreases for this group, the likelihood that an examinee guessed the correct response or misinterpreted the classification task increases.

Item-Domain and Item-Knowledge Type Divergence Indices

As previously mentioned, Rovinelli and Hambleton (1977) employed content specialists' judgments to assess the content validity of criterion-referenced test items. $I_{ik}$, an index of item-objective congruence was computed as the content specialists' average judgment that item k is a measure of objective i. Similarly in this study, the judgments of examinees were used to compute item-domain congruence, $I_{ik}^{(D)}$, and item-knowledge type congruence, $I_{ik}^{(K)}$ for each item for the domain(s) or topic(s) from which the item was generated and the type(s) of knowledge the question was asking. The professor's judgments or classifications were deemed as standard. The index of item-domain or item-knowledge type congruence was computed as the examinees' average judgment for each component of the standard judgment of the content specialist using the following formula:

$$I_{ik}^{(D)} \text{ or } I_{ik}^{(K)} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} - \sum_{\substack{m=1 \\ m\neq i}}^{N}\sum_{j=1}^{n} X_{mjk}}{2(N-1)n}$$

where,

i is the domain or type of knowledge, k is the item number;

$I_{ik}^{(D)}$ or $I_{ik}^{(K)}$ is the index of item-domain or item-knowledge type congruence for item k for the content specialists standard domain or knowledge type;

N is the number of domains or types of knowledge (i = 1,2...,N);

n is the number of examinees (j = 1,2...,n);

$X_{ijk}$ is the rating (-1 or 1) of item k as a measure of domain or knowledge type i by examinee j.

The value of 1 was assigned to an examinee's judgment if they judged an item as being generated from the same domain as judged by the content specialist. The value -1 was assigned to an examinee's judgment if they judged an item as being generated from a domain different from the content specialists. The same scheme was used for computing the index of item-knowledge type congruence. These indices range from -1 to +1 and are fairly insensitive to the number of examinees and number of domains or types of knowledge. The value of this index is subtracted from 1.0--which is the content specialist's standard judgment--to obtain $d_{ik}^{(D)}$ or $d_{ik}^{(K)}$ which refer to indices of item-domain and item-knowledge type divergence respectively:

$$d_{ik}^{(D)} = 1 - I_{ik}^{(D)} \quad , \text{and}$$

$$d_{ik}^{(K)} = 1 - I_{ik}^{(K)}$$

When the content specialist has indicated that an item has been generated from more than one domain, adjustments in the computation of $I_{ik}^{(D)}$ are made via set-theoretic logic. If two domains were deemed as the standard, then the two domains were treated as one domain thereby reducing the number of domains by one. If an examinee classified the items as being generated from only one of the two domains indicated by the content specialist, then the examinee's judgment was assigned a value of 0 rather than 1 or -1. They have classified the item partly similar and 0 is half way between 1 and -1. If an examinee classified an item as being

generated from a domain that was not judged to be the standard domain by the content specialist, no additional adjustment is needed for computing $I_{ik}^{(D)}$ since non-standard judgments are included in the computation of $I_{ik}^{(D)}$ to reduce this index regardless of whether one or more domains have been judged by the content specialist as having generated that item. If the content specialist has classified an item as being generated from three domains, then examinees' classifications were assigned a 0 if they indicated that the item was generated from one of the standard set of three domains. If the examinees' classification of an item contained two out of the three or all three of the standard set of three domains, then their classification was assigned a 1. As stated above, extraneous domains included in the examinees' classifications affect the value of $I_{ik}^{(D)}$ appropriately by reducing the magnitude of this index.

Inference Confirmation Using Interview Data

In order to confirm inferences about the items' validity, it was necessary to interview examinees about their perceptions of the meaning of each item. The thirteen examinees who volunteered were interviewed during the week immediately after the examination using a semi-structured, clinical format. All interviewees agreed to tape recording the interviews. Items deemed ambiguous were ascertained by asking examinees to state how they conceptualized the test questions. The interview began with the interviewer asking the examinee to discuss any item that contained some problem. They were not asked to discuss the first test question first, the second test question second, etc. unless, of course, they absolutely had no test question to discuss in particular from their experience with the examination. Every comment initiated by an interviewed examinee was

followed up with another question to elicit more in-depth perceptions about that particular test question until the examinee began reiterating for a second time what they had previously stated about the test question. The examinee was never interrupted while responding since their remarks were decisive for the direction and form that subsequent questions took. Because of this, questions were not known a priori. This is the basic feature that constitutes a flexible interview protocol. Pines, Novak, Posner and Vankirk (1978) stated, "The flexible interview does not enable simple discrimination, but rather provides information about what a subject is thinking, what he/she knows and believes, and why" (p. 9).

The interview data collected from self-selected examinees provided for rich descriptions of the meaning of items. As problems in items were reported by interviewees, they served as scenarios to which subsequent interviewees could agree or disagree with when no new perceptions were emanating from interviewees about those items' problems. Therefore, perceptions of problems in item meaning previously stated by other interviewees sometimes served to enhance the credibility of individual accounts of the meaning of items when conceptualizations were very similar.

The descriptions contain problems in the conceptualization of the test questions with regard to the subject matter measured by the questions, based on either examinee perceptions at the time of testing or at the time of interview. This information combined with the statistical information merges as a comprehensive qualitative and quantitative account of the examinee perceptions of the item's meanings. If the hypotheses about the meaning of the items referred from the statistical information converges with the interview data, then the procedures for utilizing the statistical information are more viable. At attempt was made to compare the classification and response data to the interview data for the interviewed examinees. In the following chapter both types of information are presented adjacently for each item rather than going through the items twice.

## RESULTS

Because of the exploratory nature of this study, Tables 8 - 28 in Appendix E[1] along with the interview data and item information discussion accompanying the table for each item do not provide for a systematic procedure for detecting items deemed anamalous by examinees. For this reason, tables for a few representative items for which some ambiguity has been reported by interviewed examinees will be discussed in detail.

The purpose here is to describe special procedures which are not as yet definitive enough to be treated as a "cut and dried" method. Therefore, the presentation of examples seem appropriate for illustrating and advocating the usefulness of this exploratory test technique. Following these examples, a summary of relevant statistical information needed to make inferences about item validity for all items is presented in Table 7 which is accompanied by a set of guidelines for using this table.

Example 1 - Item 6

> Item 6. What then is the main advantage of a parallel circuit over a series circuit? (2 pts.)

This item generated a considerable amount of discussion with the interviewed examinees. Most of the discussion revolved around the expression "main advantage." The interview data that follows are presented separately for each of eleven interviewed examinees along with whether or not they received full credit for their response as well as the topics they listed which generated the item.

Examinee 1 -

> I answered the question by interpreting the main advantage as Professor _____ thought. However, I felt that

1. Because of the already lengthy paper , Appendix E is not included in this version.

there were a couple of important things. I usually interpret these questions as what they think is the main advantage. This is a problem for me with other tests.

This examinee received full credit and classified the item as having been generated from the topics series circuit and parallel circuit.

Examinee 2 -

That was probably ambiguous. I got around it though by listing every advantage. In class, we weren't directed to one main advantage. I took two answers and incorporated them into one answer. I'm still not positive as to whic was the main advantage. The answer would probably have to be more than a negative aspect of a series circuit.

This examinee received full credit and classified the item as having been generated from topics series circuit and parallel circuit. In addition, this examinee received a perfect test score.

Examinee 3 -

I got that right away. I'd be willing to bet I was right. The main advantage is that there is the same voltage on each load.

This examinee received full credit and classified the item as having been generated from the topics series circuit and parallel circuit.

Examinee 4 -

I don't remember a main advantage mentioned. However, the answer to the question was real obvious.

This examinee received full credit and classified the item as having been generated from the topic parallel circuit. Their response was that "loads can be grouped and the disconnection of one load will not cut the circuit."

Examinee 5 -

The answer I put down was that you don't lose voltage through the circuit and all lights burn at the same brightness.

This examinee received full credit and designated the topics series circuit and parallel circuit as having generated the item.

Examinee 6 -

> This question was a direct outgrowth of the two previous questions - the characteristics of series and parallel circuits, respectively in terms of current, voltage and resistance. Voltage was constant in a parallel circuit. I was totally sure of what the main advantage was. Another advantage was that one could tamper with components without affecting the other components. If an examinee was thinking of stereo systems, however, there would be one main advantage - power.

This examinee received full credit and classified the item as having only been generated from the topic parallel circuit.

Examinee 7 -

> This question was straightforward. But for some examiness it might have been a little ambiguous. If you know the material, you know what the test maker is looking for.

This examinee received full credit and classified the item as having been generated from the topics series circuit and parallel circuit.

Examinee 8 -

> I'm not sure but it looked like the question pertained to answers in questions four and five-characteristics of a series and parallel circuit. I was trying to figure out what they were looking for. I don't recall it being stated in lecture that, "This is the main advantage."

This examinee lost one point for his response to the question. He classified the item as having been generated from the topics electric current, series circuit, and parallel circuit. This examinee's answer read, "If one element in the parallel circuit fails to function or burns out, the rest of the loads can still operate, but if a load or element in the

series circuit burns out, the whole circuit will no longer function until that element is replaced or fixed."

Examinee 9 -

> Questions like this always bother me. I know Professor _____ stressed the main advantage. I think I put down independence or when one goes out they don't all go out.

This examinee received full credit and classified the item as having been generated from the topics series circuit and parallel circuit. This examinee's response to the question actually read, "The loads are independent, and can be controlled or removed without others affected."

Examinee 10 -

> To me the main advantage of a parallel over a series circuit is that something can go out and it doesn't affect the... I can see now that in a parallel circuit you have the same voltage going to each load. I don't think I would take off a total of two points because that answer is correct. Now, I'm doubting whether I was right. The answer could be either reason.

This examinee received full credit and classified the item as having been generated from the topics series circuit and parallel circuit. In their response to this question, this examinee happened to have put down both reasons. Their total test score was fifty-one out of fifty-four.

Examinee 11 -

> People shouldn't lose credit for putting down a disadvantage of a series circuit since this implies an advantage of a parallel circuit. I thought about using the Christmas tree example, but I didn't.

This examinee received full credit for their response and classified the item as having only been generated from the topic parallel circuit.

The statistical information collected from response and classification data for Item 6 are presented in Table 3 below. In addition to the statistical information described on pages 33 and 34, z scores for the mean total test scores of cells A, B, C and D are computed separately and reported. They represent the deviation of the cell means normalized by the standard deviation of means or standard error of the mean for all scores. Or

$$z = \frac{\bar{X}_{tot_{ij}} - \bar{X}_{tot}}{s_{tot}} \sqrt{n_{ij}}$$

This additional information provides for a common metric for determining the "likelihoods" of the individual cells both within and across items.

TABLE 3

Item 6 Data

| Standard Domain | n | $d_{ik}(D)$ | Standard Knowledge Type | n | $d_{ik}(K1)$ | $d_{ik}(K2)$ |
|---|---|---|---|---|---|---|
| F. Series Circuit | 74 | 0.565 | fact | 43 | 1.047 | |
| G. Parallel Circuit | 74 | 0.113 | principle | 43 | 0.798 | |
| | | | interpretation | 43 | 0.767 | |
| Combined | 74 | 0.318 | combined | 43 | 0.419 | 0.558 |

(a) Second Credit Point

F. Series Circuit

| $1 - (\phi/\phi_{max})2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.982 | same = 0.711 | 38 | same = 0.482 |
| | diff. = 0.778 | 36 | diff. = 0.108 |
| | comb. = 0.743 | 74 | comb. = 0.340 |

Mean Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78 \quad s^2 = 26.10 \quad n = 74$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}_2$ | 42.27 | 46.82 | 45.38 | 46.21 |
| $s^2$ | 20.82 | 31.39 | 23.13 | 20.32 |
| n | 11 | 27 | 8 | 28 |
| z | -2.28 | 1.06 | -0.22 | 0.45 |

G. Parallel Circuit

| $1 - (\phi/\phi_{max})2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.678 | same = 0.723 | 65 | same = 0.337 |
| | diff. = 0.889 | 9 | diff. = 0.040 |
| | comb. = 0.743 | 74 | comb. = .340 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78 \quad s^2 = 26.10 \quad n = 74$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}_2$ | 43.33 | 46.30 | 48.00 | 47.75 |
| $s^2$ | 23.18 | 27.61 | 0 | 12.21 |
| n | 18 | 47 | 1 | 8 |
| z | -2.03 | 0.70 | 0.43 | 1.09 |

(b) First Credit Point

F. Series Circuit

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.987 | same = 0.974 | 38 | same = 0.380 |
| | diff. = 1.000 | 36 | diff. = uncomputable |
| | comb. = 0.986 | 74 | comb. = 0.058 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78 \quad s^2 = 26.10 \quad n = 74$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}_2$ | 45.00 | 45.51 | - | 46.03 |
| $s^2$ | 0 | 32.92 | - | 20.43 |
| n | 1 | 37 | 0 | 36 |
| z | -0.07 | -0.32 | - | 0.29 |

G. Parallel Circuit

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.998 | same = 0.985 | 65 | same = 0.036 |
| | diff. = 1.000 | 9 | diff. = uncomputable |
| | comb. = 0.986 | 74 | comb. = 0.058 |

Mean, Variance and n of Total Test Scores

$X_{tot} = 45.78 \quad s^2 = 26.10 \quad n = 74$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}_2$ | 45.00 | 45.48 | - | 47.78 |
| $s^2$ | 0 | 28.22 | - | 10.69 |
| n | 1 | 64 | 0 | 9 |
| z | -0.07 | -0.47 | - | 1.17 |

The considerable amount of discussion for this item expressed by interviewed examinees is reflected in the statistical information presented in Table 3 above. There was considerably higher divergence for the standard domain, series circuit, as compared to the standard domain, parallel circuit ($d_{ik}^{(D)} = 0.565$ and $d_{ik}^{(D)} = 0.113$, respectively). This indicates that there were a greater number of examinees who thought that this item was generated from "parallel circuit" than "series circuit." Since the combined $d_{ik}^{(D)} = 0.318$, there seems to be considerable agreement about the intended topic the question was supposed to be about. However, the relatively low value of the combined $d_{ik}^{(D)}$ may be misleading since the expressions "series circuit" and "parallel circuit" both appear in the actual wording of this question. Some examinees who classified the item as having been generated from both series circuit and parallel circuit, as did the professor, may have done so from mere recognition of these expressions in the question. With respect to the type(s) of knowledge asked by the question, examinees on the average disagreed more that the question required knowledge of a fact than a principle or an interpretation. For the standard set or combined knowledge types, the teaching assistant had judged the question to be asking, there was low to moderate divergence, $d_{ik}^{(K1)} = 0.419$ and $d_{ik}^{(K2)} = 0.558$.

A comparison of $1 - (\phi/\phi_{max})^2$ for the domains, series circuit and parallel circuit, for the second credit point is an indication that examinees who classified the item as having been generated only from the domain, parallel circuit, instead of both series circuit and parallel circuit, were more

likely to have received full credit for their response. In light of the higher correlation between classification and response data for the second credit point for the domain, parallel circuit, there is reason to suspect some lack of validity for this item especially since the components of the standard domain consisted of both series circuit and parallel circuit. ($1 - (\phi/\phi_{max})^2 = 0.982$ for series circuit and $1 - (\phi/\phi_{max})^2 = 0.678$ for parallel circuit.)

An inspection of the proportion of examinees receiving full credit (the second credit point) shows little difference in the proportion correct between those including and not including the domain series circuit in their classification of the item (same = 0.711 and diff. = 0.778). However the biserial correlations point to the fact that the item was more discriminating for those examinees who included series circuit in their classification. This is reflected in the relatively high mean total test score for those responding incorrectly and not including series circuit in their classification ($\bar{X}_{tot,C} = 45.38$) as compared to the mean total test scores for those responding incorrectly and classifying the item as having been generated from the domain, series circuit ($\bar{X}_{tot,A} = 42.27$). For the 8 examinees who responded incorrectly and classified differently for the second credit point for series circuit, $\bar{X}_{tot,C} = 45.38$ as compared to $\bar{X}_{tot} = 45.78$, the mean total test score for all examinees. Because of this, there is a greater likelihood that some examinees responded incorrectly because of their misinterpretation of the item's meaning rather than because of low ability. The

eight examinees who did not include series circuit in their classification and did not receive the second credit point had a mean total test score close to the mean total test score for all all examinees ($z_C$ = -0.22). These examinees, without exception, had answers that were judged equivalent to a disadvantage of a series circuit while the correct answer was an advantage of a parallel circuit. They seem to have thought that the question only pertained to a parallel circuit, yet their responses were either about parallel circuits but negatively worded or about the main disadvantage of a series circuit. An example of one of these eight responses which did not receive full credit was for a student whose total test score was 52 out of 54. The responses read, "If one appliance (for example a light bulb) goes out, the whole circuit will not shut down interrupting service."

For this item generated from two domains, the value of $d_{ik}^{(D)}$ which do not depend on the responses did not help in illuminating the ambiguous nature of the expression "main advantage" in the wording of the question as was previously discussed. It was necessary to inspect the values of $1 - (\phi/\phi_{max})^2$ for both domains at first to determine that the relationship between classification and response differed for the second credit point for the domains. A lack of validity was determined after inspection of the mean, variance, n and z score for cell C examinees.

Example 2 - Item 14

Item 14. Name the main advantage of the following (4 pts.):

A. Fuse

B. Circuit Breaker

This item generated a considerable amount of discussion with the interviewed examinees. Most of the discussion as with the sixth item revolved around the expression "main advantage." For this item, as with the sixth item, the interview data are presented separately for each of eight interviewed examinees along with whether or not they received full credit and the topic(s) they listed which generated the item.

Examinee 1 -

> There is more than one main advantage for both a fuse and a circuit breaker.

This examinee received full credit and classified the item as having been generated from the topic, abnormal circuit conditions. The topic abnormal circuit conditions was the topic that the professor had indicated the item had been generated from.

Examinee 2 -

> A guest speaker lectured to us about the main advantage of each of these. He made it very explicit. A fuse blows quickly and is more reliable whereas a circuit breaker is resetable.

This examinee received full credit for both parts of the item and classified the item as having been generated from the topics abnormal circuit conditions, ground-fault protection of equipment, ground-fault protection of people and lightning protection.

Examinee 3 -

> The expression "main advantage" was confusing, I started to write down, "In my opinion the main advantage of a fuse is ... ." I felt I know two main advantages for both fuse and circuit breakers. A fuse is more dependable and is current limiting. Then I put an arrow next to the one I thought was the main advantage. As an aside, this question appeared on an old test.

This examinee received full credit for both parts of this item and did not classify the item at the time of the test. In the lab session on the day after the test, she classified the item as having been generated from the topics electrical power system, electric current, and abnormal circuit conditions.

Examinee 4 -

> I put down current limiting for the answer to the main advantage for a fuse. Our teaching assistant talked about this question in our laboratory section today and he changed the grade for a few of us in order that we receive the full two points credit. The teaching assistnat said that the answer was it protects things better--because it is current limiting.

As this examinee explained, the answer current limiting as the main advantage of a fuse was initally marked incorrect. The examinee classified the item as having been generated from the "standard" topic, abnormal circuit conditions.

Examinee 5 -

> A fuse is current limiting or more reliable. A circuit breaker is convenient since it is resetable or reusable. The correct answers could have been more technical. However, Professor _____ wants us to know the basics because there's so much information to know in such a short period of time.

This examinee lost one of two points credit for the main advantage of a fuse and received both credit points for the main advantage of a circuit breaker. His response was "efficiency in performance, current limiting." The teaching assistant who graded the test marked his response "-1" and circled the word "main". The examinee classified the item as having been generated from the topic ground-fault protection of equipment. Because he

mentioned that the correct answer to the question could have been more technical, he rated the types of knowledge the question was asking as definition, fact and interpretation. The teaching assistant's "standard" types of knowledge the question was asking were fact and principle. This examinee may have conceptualized the item more technically than was intended as well as having been generated from a different topic than was thought by the professor as content specialist.

Examinee 6 -

> A fuse is current limiting, therefore it provides better protection to things connected to circuits. Since it is current limiting, it protects. I don't know what they were looking for current limiting or protection. Therefore, for fuse, there could be two main advantages.

This examinee received full credit for both parts of their answer to this item. His answer for the main advantage of a fuse was, "a fuse is current limiting therefore it gives maximum protection. This examinee classified this item as having been generated from the topic, lightning protection which was not the "standard" classification."

Examinee 7 -

> I didn't have one main advantage in mind. There wasn't only one main advantage. I thought the main advantage of a fuse was that it cuts the circuit more quickly than a circuit breaker which allows some voltage to go through the circuit. Everyone should have the same conception of the meaning of the words "main advantage"--the guest lecturer, teacher and test taker. Usually I think of the answer they want. That's the way I take many tests where it involves interpretation. And I compare what I think they want with how I interpret the question.

This examinee received full credit for both parts of this question and classified it as having been generated from the topic, ground-fault

protection of equipment. In addition, this examinee rated the types of knowledge the question was asking as definition, fact and principle.

Examinee 8 -

> The fuse and circuit breaker perform the same function. Something with amperage. When it works, it stops amperage right away. The circuit breaker does not and is reusable. I think the main advantage of a fuse and circuit breaker are the same.

This examinee seems to have taken a more general approach to answering this question. The main advantage according to him is tied to the general purpose of a fuse and circuit breaker rather than to the main advantage of each in contrast to the other. This examinee lost two points credit in his response to the main advantage of a fuse and classified this item as having been generated from the topics, abnormal circuit conditions and ground-fault protection of equipment.

The statistical information collected from response and classification data for Item 14 are presented in Table 4 below. As with Item 6, z scores for the mean total test scores of cells A, B, C, and D are computed separately and reported.

TABLE 4

Item 14 Data

| Standard Domain | n | $d_{ik}^{(D)}$ | Standard Knowledge Type | n | $d_{ik}^{(K1)}$ | $d_{ik}^{(K2)}$ |
|---|---|---|---|---|---|---|
| J. Abnormal Circuit Conditions | 69 | 0.394 | fact | 43 | 0.822 | |
| | | | principle | 43 | 1.066 | |
| | | | combined | 43 | 0.965 | 1.116 |

(a) Fuse - Second Credit Point

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.998 | same = 0.750 | 44 | same = 0.413 |
| | diff. = 0.720 | 25 | diff. = 0.082 |
| | comb. = 0.739 | 69 | comb. = 0.268 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot}$ = 45.78 $s^2$ = 26.76 n = 69

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 43.09 | 46.39 | 45.57 | 46.39 |
| $s^2$ | 26.49 | 18.87 | 22.29 | 42.25 |
| n | 11 | 33 | 7 | 18 |
| z | -1.72 | 0.68 | -0.11 | 0.50 |

First Credit Point

| $1 - (\phi/\phi_{max})$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.994 | same = 0.773 | 44 | same = 0.314 |
| | diff. = 0.720 | 25 | diff. = 0.057 |
| | comb. = 0.754 | 69 | comb. = 0.145 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot}$ = 45.78 $s^2$ = 26.76 n = 69

TABLE 4 (cont.)

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 43.60 | 46.15 | 46.57 | 46.00 |
| $s^2$ | 26.27 | 20.37 | 6.95 | 47.77 |
| n | 10 | 34 | 7 | 18 |
| z | -1.33 | 0.42 | 0.40 | 0.18 |

(b) Circuit Breaker - Second Credit Point

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.983 | same = 0.955 | 44 | same = 0.409 |
| | diff. = 1.000 | 25 | diff. = uncomputable |
| | comb. = 0.971 | 69 | comb. = 0.363 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 69$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 41.50 | 45.76 | - | 46.16 |
| $s^2$ | 24.50 | 21.94 | - | 35.64 |
| n | 2 | 42 | 0 | 25 |
| z | -1.17 | -0.03 | - | 0.37 |

First Credit Point

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| cannot be computed | same = 1.000 | 44 | uncomputable |
| | diff. = 1.000 | 25 | uncomputable |
| | comb. = 1.000 | 69 | uncomputable |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 6$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | - | 45.57 | - | 46.16 |
| $s^2$ | - | 22.30 | - | 35.64 |
| n | 0 | 44 | 0 | 25 |
| z | - | -0.27 | - | 0.37 |

From discussions of this item with interviewed examinees, it can be inferred that there did exist some ambiguity about the intended meaning of the expression "main advantage" particularly for the first part of the question concerning the main advantage of a fuse. For the sole domain deemed as standard by the professor, abnormal circuit conditions, $d_{ik}^{(D)} = 0.394$. This is relatively low especially in light of the fact that no part of the expression "abnormal circuit conditions" was contained in the question ensuring that the classification of the item was not a trivial exercise in recognition as might have been the case for some examinees in their classification of Item 6. With respect to the types of knowledge asked by the question, examinees on the average disagreed with both components of the standard knowledge types; $d_{ik}^{(K1)} = 0.822$ for fact and $d_{ik}^{(K1)} = 1.066$ for principle. For the standard set or combined knowledge types the teaching assistant had judged the question to be asking, there was relatively high divergence, $d_{ik}^{(K1)} = 0.965$ and $d_{ik}^{(K2)} = 1.116$.

An inspection of $1 - (\phi/\phi_{max})^2$ for the standard domain, abnormal circuit conditions, for the first part of the question concerning the main advantage of a fuse, classification and response data were considerably uncorrelated. ($1 - (\phi/\phi_{max})^2 = 0.994$ for the first credit point and $1 - (\phi/\phi_{max})^2 = 0.998$ for the second credit point.) Therefore, knowing that the question was generated from the topic, abnormal circuit conditions, was not associated with responding correctly to the question. Statistical evidence to support the misinterpretation of this item as expressed by interviewed examinees is provided by the mean total test scores of the seven

examinees who did not receive the second credit point and did not include the topic, abnormal circuit conditions, in their classification of this item. For this incorrect-different group for the part of the question about the main advantage of a fuse, $\bar{X}_{tot,C} = 45.57$ while $\bar{X}_{tot} = 45.78$ for all 69 examinees classifying this item. This produced $z_C = -0.11$ which is high enough to suggest the likelihood of misinterpretation. Even more glaring is the fact that for the seven examinees who did not receive the first credit point for the part of the question concerning the main advantage of a fuse and did not include abnormal circuit conditions in their classification, $\bar{X}_{tot,C} = 46.57$ and $z_C = 0.40$. This is the third highest mean total test score for cell C examinees across all items. In addition, the biserial correlations for those not including abnormal circuit conditions in their classification were considerably lower than for those including the topic in their classification for both credit points (see Table 4 above). For the part of the question about the main advantage of a circuit breaker, there were no examinees for either credit point who responded incorrectly and did not include abnormal circuit conditions in their classification.

Not unlike Item 6, the value of $d_{ik}^{(D)}$ did not aid in illuminating the ambiguity associated with the expression "main advantage." The values of $1 - (\phi/\phi_{max})^2$ for both credit points for the part of the question about the main advantage of a fuse were high and approximately equal. These values were not as useful as for Item 6 where the values of $1 - (\phi/\phi_{max})^2$ differed across domains. Item 14 was supposedly generated from only one

domain. As with Item 6, a lack of validity was determined after inspection of the mean, variance, n and z scores for cell C examinees. As an aside, the fact that the teaching assistant later changed some grades for this question by giving credit to the answer, current limiting, makes claims of ambiguity even more credible.

Example 3 - Item 15

> Item 15. Name the two conditions which cause a fuse to blow or a circuit to trip and briefly explain each. (4 pts.)

Aside from one examinee commenting that this question required the memorization of two facts, there was only one examinee who discussed how she interpreted this question. She stated:

> I realize now that I interpreted this question the wrong way. I took the question as wanting two examples and not two conditions to cause a fuse to blow or a circuit to trip. The two examples I wrote were plugging a large appliance into an outlet that's not meant for it, and too many systems in one outlet. I'm sure now that the correct answers were short circuit and overload. In fact, I later remembered these two conditions of overcurrent discussed in the lecture.

This examinee classified the item in the same way as the professor--as having been generated from the topic, abnormal circuit conditions. She classified the types of knowledge the item was asking as principle while the teaching assistant classified the types of knowledge the question was asking as fact and principle.

The above examinee was marked incorrect and rightly so. Giving examples instead of conditions is equivalent to not answering the question. One examinee who was not interveiwed and received a total test score of 49

out of 54 lost both credit points for his response for the second condition that causes a fuse to blow or a circuit to trip. He responded using more technical language than the other examinees. His answer read "overcurrent surge--leak of high current into system with less ampacity." The teaching assistant who graded the test wrote across the words "overcurrent surge", the word "overload." This examinee may have interpreted the word "condition" more broadly to include "problems" but not actual examples as did the interviewed examinee. This conjecture is more probable in light of this examinee's classification which consisted solely of the domain, ground-fault protection of equipment and not, abnormal circuit conditions--the standard domain as shown in Table 5 below.

TABLE 5

Item 15 Data

| Standard Domain | n | $d_{ik}^{(D)}$ | Standard Knowledge Type | n | $d_{ik}^{(K1)}$ | $d_{ik}^{(K2)}$ |
|---|---|---|---|---|---|---|
| J. Abnormal Circuit Conditions | 69 | 0.252 | fact | 43 | 0.667 | |
| | | | principle | 43 | 0.884 | |
| | | | combined | 43 | 0.663 | 0.969 |

(a) Condition One - Second Credit Point

| $1 - (\phi/\phi_{max})$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.983 | same = 0.887 | 53 | same = 0.241 |
| | diff. = 0.813 | 16 | diff. = 0.013 |
| | comb. = 0.870 | 69 | comb. = 0.146 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 69$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 43.83 | 46.00 | 46.00 | 45.85 |
| $s^2$ | 38.17 | 20.35 | 3.00 | 55.14 |
| n | 6 | 47 | 3 | 13 |
| z | -0.92 | 0.29 | 0.07 | 0.05 |

First Credit Point

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.878 | same = 0.981 | 53 | same = 0.935 |
| | diff. = 0.938 | 16 | diff. = 0.143 |
| | comb. = 0.971 | 69 | comb. = 0.532 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 69$

TABLE 5 (cont.)

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 35.00 | 45.96 | 44.00 | 46.00 |
| $s^2$ | 0 | 20.27 | 0 | 47.43 |
| n | 1 | 52 | 1 | 15 |
| z | -2.08 | 0.25 | -0.34 | 0.16 |

(b) Condition Two - Second Credit Point

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.999 | same = 0.943 | 53 | same = -0.026 |
| | diff. = 0.938 | 16 | diff. = -0.238 |
| | comb. = 0.942 | 69 | comb. = -0.093 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 69$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 46.00 | 45.74 | 49.00 | 45.67 |
| $s^2$ | 39.00 | 21.91 | 0 | 46.95 |
| n | 3 | 50 | 1 | 15 |
| z | 0.07 | -0.05 | 0.62 | -0.82 |

First Credit Point

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.878 | same = 0.981 | 53 | same = 0.587 |
| | diff. = 0.938 | 16 | diff. = 0.238 |
| | comb. = 0.971 | 69 | comb. = 0.151 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.78$ $s^2 = 26.76$ $n = 69$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 39.00 | 45.89 | 49.00 | 45.67 |
| $s^2$ | 0 | 21.67 | 0 | 46.95 |
| n | 1 | 52 | 1 | 15 |
| z | -1.31 | 0.15 | 0.62 | -0.82 |

For this item, the values of $d_{ik}^{(D)}$ and $1 - (\phi/\phi_{max})^2$ are not helpful for detecting the perceived ambiguity or over-interpretation which may have occurred for this one examinee for both credit points for the second condition which causes a fuse to blow or a circuit to trip. The mean, n and z score for the total test scores for those responding incorrectly and classifying differently (cell C) comprise the statistical information which illuminates the likelihood of misinterpretation for this one examinee. $\bar{X}_{tot,C} = 49.00$ , $n = 1$ , and $z_C = 0.62$ . As for the other statistical information, there is no other apparent evidence suggesting a lack of validity for this item.

Example 4 - Item 21

Item 21. What is the difference between electro-motive force (EMF) and a voltage drop? (2 pts.)

This item provided for the greatest amount of discussion. It was likely the most confusing and/or ambiguous as expressed by the interviewed examinees. Most of the discussion revolved around the word "difference." The "standard" topics from which the item was deemed to have been generated from were electromotive force and circuit equations. The "standard" types of knowledge this question asked was definition. For this item the interview data are presented separately for each of the interviewed examinees along with how many points they received from the two credit points possible, the topics they listed as generating the item and the types of knowledge they felt the question asked.

Examinee 1 -

> I think what I put down was right. Others seemed confused about that question after the test when we spoke about it.

This examinee received the full two credit points for her response to this item. Her answer read, "Electromotive force occurs only at the source, while a voltage drop occurs at elements not at the source." She classified the item as having been generated from the topic, electromotive force and judged the type of knowledge the question was asking in the "standard" way as only definition.

Examinee 2 -

> This item was kind of confusing with regard to the expression "the difference between EMF and voltage drop." They're related, but they are entirely different things. I wasn't sure of the context in which the question was asked. The question should have defined exactly what they're looking for.

This examinee received the full two credit points for this item and classified it as having been generated from the topic electromotive force. He judged the types of knowledge the question was asking as definition, fact, principle and interpretation. The fact that he indicated interpretation may reflect his expressed confusion.

Examinee 3 -

> I got it right. In fact I received a perfect score on the test. If you do know the difference it's really clear. If one were to think that electromotive force and voltage drop were not comparable, then he would be reading more into the question and think there is more to the answer than there is.

This examinee received the full credit for her response. Her answer read, "EMF is what comes into the system from the source. A voltage drop occurs

in the system due to resistance." She classified the item as having been generated from the topic, electromotive force and judged the type of knowledge the question was asking as principle.

Examinee 4 -

> I remember seeing it in my notes. EMF is at the source and voltage drop at loads. I interpreted "differences" as different locations. There could be other differences, however, like input and output. EMF and voltage drop were really not in the same category. What is the difference infers that there is one thing that is the difference. One could have thought that EMF is at the source while voltage drop is in the circuit.

This examinee received full credit for their response. He classified the item as having been generated from the topic, electric current and judged the type of knowledge the question asked as fact.

Examinee 5 -

> What they basically wanted was the difference between electromotive force and voltage drop. It was a very straightforward question. EMF is what goes into the system at the very beginning and voltage drop is what occurs at each resistance. The first time I read the question I said to myself one is one and one is the other. They wanted a definition of one and the definition of the other. The two definitions would include the difference between the two.

This examinee received full credit. She classified the item as having been generated from the topic, electromotive force and judged the types of knowledge the question was asking as definition, principle, and interpretation.

Examinee 6 -

> That was a fair question mentioned in the syllabus and in the lecture. EMF is a potential for flow of charges and voltage drop is how much that electron will lose going through a load. Voltage drops will add to the EMF. It's a technical difference.

> EMF is more of a concept. However, I think they're related since both are measured in the same terms. One is measured at the source and one's measured in the circuit.

This examinee lost one of two credit points for his response. His answer read, "EMF is the potential for current flow whereas voltage drops are the amount of energy lost by the charges after having passed through a load." He classified the item as having been generated from the topic, electromotive force and judged the types of knowledge the question was asking as definition and interpretation. The fact that this examinee expressed that the difference was technical might be related to his indicating interpretation as a type of knowledge the question was asking. As with all but one examinee who was interviewed, this examinee did not classify the item as having been generated from the "standard" topic, circuit equations.

Examinee 7 -

> I knew what EMF was and I knew what voltage drop was. I didn't at all see how they related. The question didn't seem clear to me. Again, it could have been my lack of knowledge. I knew what EMF was and I knew what voltage drop was--but I didn't see how anyone couldn't be confused because they were the same type of thing, I don't know. So I don't know why the question was asked. I went ahead and defined each thing and hoped that would show the difference.

The examinee, received both credit points for his response. His answer read, "EMF is a source of power and voltage drop occurs at individual loads along circuit." He classified the item as having been generated from the topics, electromotive force, electric current and resistance. They judged the types of knowledge the question was asking as definition, fact and interpretation.

Examinee 8 -

> I don't know if I answered it correctly. The two things, electromotive force and voltage drop seemed unrelated. EMF is a source of energy while voltage drop is losing voltage in the system. I looked at the two expressions as concepts in themselves. The word "difference" is crucial. After talking about it, it is not the same as taking the question on a test.

This examinee received the full two points credit. His response read, "EMF is the voltage produced by a source and voltage drop is the voltage loss due to transporting the energy." This examinee classified the item as having been generated from the topic, electrical power system and judged the types of knowledge the question was about as definition and principle.

Examinee 9 -

> You had to think about this question. You had to think about how the system works. The word "difference" is what could have made the question confusing. EMF is the initial source which drops around the system. One is more qualitative than the other. EMF is a push which starts the system. On the other hand, voltage drop is a process. When you see difference, you categorize things together like when you compare "black" to "wall". Black is a color and wall is part of a building. That's the way I think of difference.

This examinee received both credit points for this item. His answer read, "An EMF is what the source produces and the voltage drop is the difference of the EMF leaving the source and the remaining voltage coming back to the source (creating the circuit)." This examinee did not classify this item as having been generated from any special topic. He classified all the other items however. He judged the types of knowledge the question was asking as definition, principle and interpretation.

Examinee 10 -

> For this question about voltage drop and EMF, you can go forever on differences. It would have been better if they asked for the definitions of the two terms separately. It would have been more straightforward. EMF is input and the total voltage drop.

This examinee received the full two points credit. He classified the item as having been generated from the topics, electromotive force and circuit equations. He judged the types of knowledge the question was asking as definition and interpretation.

From the statistical information collected from the response and classification data presented in Table 6 below, it is immediately apparent from looking at the divergence measures for this item, that for the domain, circuit equations, $d_{ik}^{(D)} = 1.068$. This is the highest value of $d_{ik}^{(D)}$ across all items. In fact, only 1 out of 68 examinees who classified this item included the topic, circuit equations, in their classification.

TABLE 6

Item 21 Data

| Standard Domain | n | $d_{ik}^{(D)}$ | Standard Knowledge Type | n | $d_{ik}^{(K1)}$ | $d_{ik}^{(K2)}$ |
|---|---|---|---|---|---|---|
| B. Electromotive Force | 68 | 0.244 | definition | 43 | 0.775 | 0.775 |
| E. Circuit Equations | 68 | 1.068 | | | | |
| Combined | 68 | 0.609 | | | | |

(a) Second Credit Point

B. Electromotive Force

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.881 | same = 0.736 | 53 | same = 0.077 |
| | diff. = 0.667 | 15 | diff. = 1.029 |
| | comb. = 0.721 | 68 | comb. = 0.417 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.69$ $s^2 = 27.65$ $n = 68$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 45.43 | 46.00 | 36.40 | 49.50 |
| $s^2$ | 9.50 | 22.95 | 50.80 | 8.06 |
| n | 14 | 39 | 5 | 10 |
| z | -0.19 | 0.37 | -3.95 | 2.29 |

E. Circuit Equations

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.994 | same = 1.000 | 1 | same = uncomputable |
| | diff. = 0.716 | 67 | diff. = 0.419 |
| | comb. = 0.721 | 68 | comb. = 0.417 |

TABLE 6 (cont.)

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot}$ = 45.69 $s^2$ = 27.65 n = 68

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | - | 45.00 | 43.05 | 46.75 |
| $s^2$ | - | 0 | 34.83 | 22.11 |
| n | 0 | 1 | 19 | 48 |
| z | - | -0.13 | -2.19 | 1.40 |

(b) First Credit Point

B. Electromotive Force

| $1 - (\phi/\phi_{max})^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.731 | same = 0.943 | 53 | same = 0.059 |
| | diff. = 0.667 | 15 | diff. = 1.029 |
| | comb. = 0.882 | 68 | comb. = 0.674 |

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot}$ = 45.69 $s^2$ = 27.65 n = 68

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | 45.33 | 45.88 | 36.40 | 49.50 |
| $s^2$ | 1.33 | 20.31 | 50.80 | 8.06 |
| n | 3 | 50 | 5 | 10 |
| z | -0.12 | 0.26 | -3.95 | 2.29 |

E. Circuit Equations

| $1 - \phi^2$ | Proportion Correct | n | $r_{biserial}$ |
|---|---|---|---|
| 0.998 | same = 1.000 | 1 | same = uncomputable |
| | diff. = 0.881 | 67 | diff. = 0.673 |
| | comb. = 0.882 | 68 | comb. = 0.674 |

TABLE 6 (cont.)

Mean, Variance and n of Total Test Scores

$\bar{X}_{tot} = 45.69$ $s^2 = 27.65$ $n = 68$

| | incorrect - same | correct - same | incorrect - diff. | correct - diff. |
|---|---|---|---|---|
| $\bar{X}$ | - | 45.00 | 39.75 | 46.51 |
| $s^2$ | - | 0 | 50.79 | 20.25 |
| n | 0 | 1 | 8 | 59 |
| z | - | -0.13 | -3.20 | 1.20 |

From the large amount of discussion about this item with the interviewed examinees, this very high value of $d_{ik}^{(D)}$ should not be viewed artifactually. The large differences between the biserial correlations for the groups of examinees who did not include the topics, electromotive force and circuit equations, in their classification respectively, $r_b = 1.029$ and $r_b = 0.419$ in part supports the above assertion. Examinees who did not include electromotive force and responded correctly to the second credit point for this item had considerably higher total test scores than examinees who did not include circuit equations in their classification and responded correctly ($\bar{X}_{tot,D} = 49.50$ for electromotive force and $\bar{X}_{tot,D} = 46.75$ for circuit equations). Examinees who did not include electromotive force in their classification and responded incorrectly for the second credit point had considerably lower total test scores than those examinees who did not include circuit equations in their classification and responded incorrectly ($\bar{X}_{tot,C} = 36.40$ for electromotive force and $\bar{X}_{tot,C} = 43.05$ for circuit equations).

The word "difference" in the wording of the item is intuitively related to the concept of circuit equations especially in light of the fact that the difference between electromotive force and voltage can be described as a difference in volts (input vs. output). However, this does not discredit the claim of ambiguity for this item as expressed by interviewed examinees. The 19 examinees who responded incorrectly and did not include circuit equations in their classification yielded $\bar{X}_{tot,C} = 43.05$ which is not very low. The fact that for this group $s^2 = 34.83$ is indicative of there being

more than a handful of examinees who had total test scores higher than $\bar{X}_{tot}$ for all 68 examinees. This could not have been determined from an inspection of $z_C$ only ($z_C = -2.19$). The high value of $d_{ik}^{(D)}$ for the domain, circuit equations, alone justifies some lack of validity for this item in the form of misinterpretation of the intended meaning of the item. In fact, the item may have been perceived as less ambiguous if it were worded, "What is the difference in meaning between electro-motive force (EMF) and a voltage drop"?

The purpose of the above discussions of items that were deemed ambiguous by self-selected interviewed examinees serves to emphasize the exploratory state-of-the-art for the methods proposed in this study. For example, for Item 21, $d_{ik}^{(D)}$ was very useful while $z_C$ alone could have concealed problems with interpreting the meaning of the item. For Item 6, $d_{ik}^{(D)}$ was not useful while $1 - (\phi/\phi \text{ max})^2$ across domains and $z_C$ aided in illuminating problems experienced by a number of examinees. For Item 14, $z_C$ was most useful while for Item 15, $\bar{X}_{tot,C}$ and $n_C$ represented the most informative statistical information.

The variability of the statistical information used for making inferences about the lack of validity for some items coupled with the fact that the list of topics that is considered to have generated the items was not purposely constructed for detecting ambiguity, but rather sub-unit headings taken from the instructional manual accounts for the unsystematic nature of the proposed methods. Before discussing the strengths and limitations of the quantitative methods proposed in this

study and offering guidelines for their appropriate use, a summary table of important statistical information for all twenty-one items is presented in Table 7 below.

TABLE 7

Summary of Important Statistical Information Used to Make Inferences About Item Validity

| Item | Credit point | Standard Domain | $d_{ik}$ (D) | $d_{ik}$ (K1) | $d_{ik}$ (K2) | $1 - (\phi/\phi_{max})^2$ | Prop. Correct Comb. | $r_{bis.}$ comb. | $z_A$ | $z_B$ | $z_C$ | $z_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | A | 0.226 | 0.693 | 0.409 | 0.756 | 0.938 | 0.651 | -3.02 | 0.58 | -1.19 | 0.47 |
| | 2 | | | | | 0.997 | 0.988 | -0.514 | 1.34 | -0.13 | - | -0.07 |
| | 3 | | | | | 0.962 | 0.864 | 0.405 | -2.23 | 0.87 | -0.62 | 0.25 |
| | 4 | | | | | 0.993 | 0.975 | 1.171 | -3.88 | 0.73 | - | -0.07 |
| | 5 | | | | | 0.442 | 0.938 | 0.730 | -3.70 | 0.51 | -1.75 | 0.88 |
| | 6 | | | | | 0.999 | 0.889 | 0.447 | -2.82 | 0.95 | -2.59 | 1.12 |
| 2 | 7 | A | 0.423 | 0.341 | 0.288 | 0.666 | 0.921 | 0.109 | 0.31 | 0.74 | -1.91 | -0.64 |
| | | B | 0.791 | | | 0.866 | | | -1.91 | 0.74 | 0.31 | -0.28 |
| 3 | 8 | C | 0.342 | 0.705 | 0.705 | 0.994* | 0.987 | 0.082 | -0.04 | -0.55 | - | 0.83 |
| | 9 | | | | | ** | 1.000 | ** | - | -0.56 | - | 0.83 |
| 4 | 10 | F | 0.051 | 0.961 | 0.876 | 0.996* | 0.909 | 0.892 | -4.25 | 1.30 | - | 0.35 |
| | 11 | | | | | 0.975 | 0.792 | 0.467 | -2.82 | 1.35 | 0.62 | -0.01 |
| | 12 | | | | | 0.998* | 0.948 | 0.725 | -2.97 | 0.64 | - | 0.35 |
| 5 | 13 | G | 0.037 | 0.992 | 0.822 | 0.869 | 0.784 | 0.676 | -3.25 | 1.69 | -0.93 | 0.83 |
| | 14 | | | | | 0.845 | 0.824 | 0.623 | -3.17 | 1.43 | -0.93 | 0.83 |
| | 15 | | | | | 0.822 | 0.865 | 0.688 | -3.38 | 1.29 | -0.93 | 0.83 |
| 6 | 16 | F | 0.565 | 0.419 | 0.558 | 0.987* | 0.986 | 0.058 | -0.07 | -0.32 | - | 0.29 |
| | 17 | | | | | 0.982 | 0.743 | 0.340 | -2.28 | 1.06 | -0.22 | 0.45 |
| | 16 | G | 0.113 | | | 0.998* | 0.986 | 0.058 | -0.07 | -0.47 | - | 1.17 |
| | 17 | | | | | 0.678 | 0.743 | 0.340 | -2.03 | 0.70 | 0.43 | 1.09 |
| 7 | 18 | N | 0.696 | 1.081 | 1.054 | 0.984* | 0.972 | 0.738 | - | 0.50 | -2.40 | 0.13 |

TABLE 7 (cont.)

| Item | Credit Point | Standard Domain | $d_{ik}$ (D) | $d_{ik}$ (K1) | $d_{ik}$ (K2) | $1 - (\phi/\phi_{max})^2$ | Prop. Correct Comb. | $r_{bis.\ comb.}$ | $z_A$ | $z_B$ | $z_C$ | $z_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 19 | | | | | 0.916 | 0.994 | 0.649 | -1.89 | 0.90 | -1.93 | 0.13 |
| 8 | 20 | F | 0.270 | 0.802 | 0.829 | 0.966 | 0.958 | 0.019 | -0.10 | -0.12 | 0.02 | 0.31 |
| | 21 | | | | | 0.995 | 0.662 | 0.314 | -1.90 | 1.15 | 0.06 | 0.34 |
| | 20 | G | 0.255 | | | 0.961 | 0.958 | 0.019 | -1.10 | -0.16 | 0.02 | 0.20 |
| | 21 | | | | | 0.986 | 0.662 | 0.314 | -1.90 | 1.20 | 0.06 | 0.21 |
| 9 | 22 | H | 0.443 | 0.907 | 0.953 | 0.890* | 0.930 | 0.743 | - | 1.16 | -3.18 | -0.07 |
| | 21 | | | | | 0.484 | 0.915 | 0.696 | -0.53 | 1.26 | -3.18 | -0.07 |
| | 22 | A | 0.881 | | | 0.930 | 0.930 | 0.743 | -0.62 | -0.13 | -3.67 | 1.03 |
| | 23 | | | | | 0.966 | 0.915 | 0.696 | -0.62 | -0.13 | -3.39 | 1.10 |
| 10 | 24 | D | 0.757 | 1.109 | 1.054 | 0.942 | 0.662 | 0.567 | -1.26 | -0.11 | -2.73 | 2.44 |
| | 25 | | | | | 0.992 | 0.662 | 0.567 | -1.26 | -0.11 | -2.73 | 2.44 |
| | 24 | E | 0.348 | | | 0.999 | 0.662 | 0.567 | -2.32 | 2.01 | -1.92 | 0.86 |
| | 25 | | | | | 0.999 | 0.662 | 0.567 | -2.32 | 2.01 | -1.92 | 0.86 |
| 11 | 26 | E | 0.816 | 0.607 | 0.500 | 1.000 | 0.754 | 0.696 | -1.46 | 1.31 | -3.42 | 1.65 |
| | | F | 0.440 | | | 0.965 | | | -2.98 | 1.54 | -2.15 | 1.44 |
| | | G | 0.489 | | | 1.000 | | | -3.17 | 1.48 | -1.93 | 1.50 |
| | 27 | E | | | | 0.999 | 0.768 | 0.740 | -1.46 | 1.31 | -3.71 | 1.70 |
| | | F | | | | 0.981 | | | -3.24 | 1.58 | -2.15 | 1.44 |
| | | G | | | | 0.997 | | | -3.48 | 1.54 | -1.93 | 1.50 |
| | 28 | E | | | | 0.667 | 0.884 | 0.420 | 0.04 | 0.40 | -2.15 | 0.60 |
| | | F | | | | 0.904 | | | -1.47 | 0.32 | -1.44 | 0.77 |
| | | G | | | | 0.994 | | | -1.81 | 0.38 | -0.93 | 0.67 |
| | 29 | E | | | | 0.861 | 0.841 | 0.399 | -0.76 | 0.68 | -1.87 | 0.61 |

TABLE 7 (cont.)

| Item | Credit Point | Standard Domain | $d_{ik}$ (D) | $d_{ik}$ (K1) | $d_{ik}$ (K2) | $1 - (\phi/\phi_{max})^2$ | Prop. Correct Comb. | $r_{bis.}$ comb. | $z_A$ | $z_B$ | $z_C$ | $z_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | | | | 0.939 | | | -1.66 | 0.51 | -1.16 | 0.78 |
| | | G | | | | 0.993 | | | -2.26 | 0.64 | -0.51 | 0.60 |
| | 30 | E | | | | 0.999 | 0.768 | 0.455 | -1.81 | 1.51 | -1.67 | 0.61 |
| | | F | | | | 0.981 | | | -1.96 | 0.85 | -1.38 | 1.05 |
| | | G | | | | 0.975 | | | -1.45 | 0.41 | -1.93 | 1.61 |
| | 31 | E | | | | 1.000 | 0.696 | 0.596 | -2.11 | 1.86 | -2.38 | 1.27 |
| | | F | | | | 0.994 | | | -2.06 | 1.11 | -2.52 | 1.96 |
| | | G | | | | 0.986 | | | -0.71 | 1.11 | -2.07 | 1.96 |
| 12 | 32 | E | 0.948 | 0.581 | 0.558 | 0.994 | 0.543 | 0.656 | -1.32 | 1.02 | -2.33 | 2.54 |
| | | F | 0.784 | | | 0.998 | | | -1.07 | 0.74 | -2.50 | 2.75 |
| | | G | 0.322 | | | 0.950 | | | -2.26 | 2.58 | -1.41 | 0.96 |
| | 33 | E | | | | 0.999 | 0.486 | 0.572 | -1.32 | 1.02 | -2.93 | 2.77 |
| | | F | | | | 0.988 | | | -1.74 | 1.21 | -2.71 | 2.74 |
| | | G | | | | 0.942 | | | -2.91 | 2.85 | -1.47 | 0.91 |
| 13 | 34 | H | 0.222 | 1.128 | 1.124 | 1.000 | 0.855 | 0.256 | -1.52 | 0.41 | 0.20 | 0.42 |
| | 35 | | | | | 0.999 | 0.812 | 0.137 | -1.27 | 0.37 | 0.85 | 0.02 |
| 14 | 36 | J | 0.394 | 0.394 | 1.116 | 0.994 | 0.754 | 0.145 | -1.33 | 0.42 | 0.40 | 0.18 |
| | 37 | | | | | 0.998 | 0.734 | 0.268 | -1.27 | 0.68 | -0.11 | 0.50 |
| | 38 | | | | | ** | 1.000 | ** | -- | -0.27 | -- | 0.37 |
| | 39 | | | | | 0.983* | 0.971 | 0.363 | -1.17 | -0.03 | -- | 0.37 |
| 15 | 40 | J | 0.252 | 0.663 | 0.969 | 0.878 | 0.971 | 0.532 | -2.08 | 0.25 | -0.34 | 0.16 |

TABLE 7 (cont.)

| Item | Credit Point | Standard Domain | $d_{ik}(0)$ | $d_{ik}(K1)$ | $d_{ik}(K2)$ | $1 - (\phi/\phi_{max})^2$ | Prop. Correct Comb. | $r_{bis.\ comb.}$ | $z_A$ | $z_B$ | $z_C$ | $z_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 41 | | | | | 0.983 | 0.870 | 0.146 | -0.92 | 0.29 | 0.07 | 0.05 |
| | 42 | | | | | 0.878 | 0.971 | 0.151 | -1.31 | 0.15 | 0.62 | -0.82 |
| | 43 | | | | | 0.999 | 0.942 | -0.093 | 0.07 | -0.05 | 0.62 | -0.82 |
| 16 | 44 | J | 0.381 | 0.965 | 1.054 | 0.985 | 0.899 | 0.622 | -1.32 | 0.12 | -2.88 | 1.49 |
| | 45 | | | | | 0.987 | 0.565 | 0.179 | -0.24 | -0.16 | -1.12 | 1.65 |
| | 46 | | | | | 0.997 | 0.696 | 0.502 | -1.48 | 0.62 | -2.40 | 2.16 |
| | 47 | | | | | 0.987 | 0.377 | 0.383 | -0.86 | 0.62 | -1.38 | 2.61 |
| 17 | 48 | L | 0.246 | 0.872 | 0.853 | 0.870 | 0.942 | 0.375 | -0.72 | 0.19 | -1.40 | 0.46 |
| | 49 | | | | | 0.927 | 0.899 | 0.401 | -2.08 | 0.64 | -0.44 | 0.12 |
| 18 | 50 | B | 0.519 | 0.225 | 0.225 | ** | 1.000 | ** | -- | 1.52 | -- | -1.69 |
| 19 | 51 | C | 0.238 | 0.233 | 0.233 | ** | 1.000 | ** | -- | 0.77 | -- | -1.44 |
| 20 | 52 | H | 0.095 | 0.240 | 0.240 | ** | 1.000 | ** | -- | 0.01 | -- | -0.04 |
| 21 | 53 | B | 0.244 | 0.775 | 0.775 | 0.731 | 0.882 | 0.674 | -0.12 | 0.26 | -3.95 | 2.29 |
| | | E | 1.068 | | | 0.998* | | | -- | -0.13 | -3.20 | 1.20 |
| | 54 | B | | | | 0.881 | 0.721 | 0.417 | -0.19 | 0.37 | -3.95 | 2.29 |
| | | E | | | | 0.994 | | | -- | -0.13 | -2.19 | 1.40 |

* indicates that $1 - \phi^2$ was used instead of $1 - (\phi/\phi_{max})^2$.

** indicates that this value cannot be computed.

For each credit point of each item and for each component of the standard classification when applicable, the divergence measures $d_{ik}^{(D)}$, and the combined divergence measures $d_{ik}^{(K1)}$ and $d_{ik}^{(K2)}$, $1 - (\phi/\phi \text{ max})^2$ or $1 - \phi^2$, the combined proportion correct, the combined biserial correlation and the z scores for cells A,B,C and D are presented. (For the other statistical information, one must refer to Tables 8 - 28 in Appendix E where complete information is available for each item.)

Assuming that the topic headings were carefully constructed, first the high values of $d_{ik}^{(D)}$ should be noted. In cases where an item was generated from more than one domain, the domain associated with the high value of $d_{ik}^{(D)}$ should also be noted. Higher values of $d_{ik}^{(D)}$ mean that a smaller number of examinees classified the item as having been generated from that domain (overall disagreement).

The second general indicator provided in Table 7 for detecting items whose validity should be suspect is $1 - (\phi/\phi \text{ max})^2$ or $1 - \phi^2$. This index is mostly useful for items that were generated from more than one domain. If there is a considerable discrepancy for the values of $1 - (\phi/\phi \text{ max})^2$ across domains for one item, then responding correctly is more highly correlated with knowing that the item was generated from the domain with the smaller value of $1 - (\phi/\phi \text{ max})^2$. Items containing such discrepancies should also be noted. It is likely that items with high values of $d_{ik}^{(D)}$ will not have considerable discrepancies between values of $1 - (\phi/\phi \text{ max})^2$.

The most specific and perhaps most informative indicator of potential ambiguity in Table 7 is $z_C$. The higher the value of $z_C$, the higher are

the mean total test scores among those who responded incorrectly and did not include the associated standard domain component in their classification. Small negative values of $z_C$ should be noted in addition to all positive values. The usefulness of the other statistical information, especially $z_A$ will not be discussed here since it would involve adjusting these values as a function of item difficulty and perhaps examinee total classification scores.

The three indicators discussed above should have provided the test user with a number of items that are suspect with respect to their validity in the sense that examinees may have misinterpreted their meaning. Depending on which measure was used to bring attention to an item, items noted as potentially ambiguous can be re-examined for clarity using the additional information provided by the meaning of each measure. For $d_{ik}^{(D)}$, the additional information consists of the classification information without regard to response information. For discrepancies across domains for $1 - (\phi/\phi_{max})^2$, the additional information is focused on the relationship between response and classification data across all examinees. And for $z_C$, the additional information is in the form of total test scores for examinees who both responded to an item incorrectly and classified it differently. From these measures for items noted as potentially ambiguous, identification of technical defects in the wording or perceived meaning of the item can be aided. By knowing more about the potential source of misinterpretation, the inductive process of finding previously unnoticed technical defects is facilitated.

Examinee Classification Behavior and Perceptions of the Classification Task

On the item information form that was distributed in the laboratory sections on the day following the examination, examinees were asked to describe the method they used for classifying the test items. Their responses to this question revealed seven to eight different methods for the forty-four respondents who answered. The most frequently reported method was answering each item and then classifying that item before going on to the next question. Sixteen of the examinees reported using this method. Thirteen examinees reported that they had classified the items after having answered all the questions. From the comments appearing on the item information forms, this method was likely associated with examinees' fears that they might not have had enough time to finish the test or that they were generally anxious about taking tests. Four examinees stated that they answered 3 or 4 questions and then went back to classify them. One of these students remarked that in answering the questions, they would get into a "role" or "streak", and when the streak ended for one reason or another, they would classify the questions just answered. Another of these examinees reported that the reason they answered three or four questions in a row was because the questions as they perceived were often grouped by topic. Therefore, answering questions associated with topics made for a convenient way to take the test and classify the items. Four examinees reported that they did not classify during the test, but during the laboratory class just prior to answering this question. Their reasons for not

classifying during the test had to do with thinking that they did not have enough time to classify. Four examinees surprisingly claimed that they classified each item first then answered the item before going on to the next item to classify it. Two other examinees expressed that they did not classify the items in any systematic way, but rather haphazardly. And the remaining examinee stated that they classified the items on a particular page of the test after answering the items on that page of the test.

Five of the interviewed examinees commmented about the classification of items during the test. What they had to say is presented for each examinee separately.

Examinee 1 -

> The classification took up time on the exam. I was concerned about taking the exam in the first place. There may have been some inaccuracies in my classifications. If I was really concerned about the classifications, I would have tried to put a classification for every item. As it turned out, I went through the items only once to classify them.

Examinee 2 -

> Having to look at the items to classify them as to the type of knowledge each was asking on the day after the test was drudgery. Perhaps it would have been better to do these classifications a week later when we got the test back. But, classifying the items during the test had no negative effect.

This examinee was one of the four who classified each item first and then answered that item before going on the the next question. This examinee received a total test score of 49 out of a possible 54 credit points.

Examinee 3 -

I classified after I answered all the questions. I thought it was irritating to do it during the test.

Examinee 4 -

Classifying wasn't much of a burden. I didn't think I had that much time left. With the normal pressure of the test, I didn't think we needed anything else.

I started off doing the question then the classification all the way through until the last page...

I think the test covered everything. We never talked about lightning protection--I didn't find anything on the test having to do with lighting protection. I didn't study this topic and when I saw lighting protection in the list of topics, I got worried that there would be something on the test from that section. I thought there would be a question for each topic.

This examinee received a total test score of 44.

Examinee 5 -

I read the question first, figured out what part of the electricity section the question was referring to, and then put down the classification. Then I answered the question.

I put down classifications that would help me answer the question and classifications that helped me generate my own questions about the test question and this helped me arrive at the answer. In fact, because of the classification task, I'm sure that I did better on the test than if I had not classified.

This examinee received a total test score of 44.

Because of the newness of this idea, and because examinees were instructed at the beginning of the test that the classifications would not affect their grade, the reliability of the item classification scores is likely to be low. Therefore, no attempt was made in this study to obtain

examinee total classification scores and compare them or correlate them with examinee total test scores. It can be conjectured that the classification of an item right after responding to it should be more reliable since the examinees would be more aware of the domain in which they spent most mental effort. However, in light of some of the comments, the possibility exists that classifying prior to answering may enhance examinee performance.

# DISCUSSION

This study has attempted to use students' judgments of test items for making inferences about the validity of the items of one criterion-referenced achievement test on electricity. This chapter serves to characterize the methods employed for making these inferences about the validity of this test's items viewed in the interpretive sense, and offers guidelines for the improved use of the proposed methods in a number of testing contexts or situations and suggestions for further research that must occur before classification procedures can be employed non-experimentally and most beneficially in actual testing settings.

Based on the discussions in the previous chapter and item information discussions in Appendix E, a conceivable lack of validity in the interpretive sense was inferred for Items 6, 14, 15, and 21. For Items 6 and 14, expressed misinterpretation from interviewed examinees revolved around the expression "main advantage" in the wording of these items. For Item 15, it is hypothesized that misinterpretation revolved around the level of specificity or generality intended by the word "conditions." For Item 21, expressed ambiguity was attributable to the impreciseness of the word "difference" in the wording of the question. From the discussions in Appendix E, a lack of curricular validity or content relevance was inferred for Items 2, 11 and 12. For Item 2, the distinction between carbon-zinc and alkaline batteries was not seen as relevant for the study of electricity in architecture. For Items 11 and 12, some interviewed examinees claimed

that these questions required knowledge of lighting and bulblife from the previous unit on illumination in combination with the present electricity unit. Hence, these parts of these questions were seen by some as unfair. However, it may have been one of the course objectives to require knowledge of illumination for the electricity test. This would disconfirm the curricular invalidity claims for Items 11 and 12. And the true-false question about batteries could have justifiably been included in the test to measure examinees' general awareness of electricity and therefore be considered curricularly valid. The main focus of this research was to detect item ambiguity as reflected in the perceived meaning of the items by examinees. Examinee classification of items thereby served as a "reflective dimension" (Proppé, 1979).

Items 6, 14 and 21 were perceived as possibly ambiguous by the interviewed examinees. Item 15 interview data did not illuminate the possibility of ambiguity connected to the perceived meaning of the word "conditions" for this item. The statistical information combined with an error analysis of the responses to the item shed light on this possibility. Assuming that the "standard" domain labels or topic headings are well-constructed (to be discussed later in this chapter), the most important statistical measures for assessing the likelihood that examinees' perceived meaning of an item lacked validity in the interpretive sense are $d_{ik}^{(D)}$ for individual components of the standard classification (as with Item 21); discrepancies between $1 - (\phi/\phi_{max})^2$ across domains for items generated from more than one domain (as with Item 6); $X_{tot,C}$, $s^2_{tot,C}$, $n_C$ and $z_C$ (as with Item 14); and $X_{tot,C}$ and $n_C$ (as with Item 15). High values of $d_{ik}^{(D)}$ or high

item-domain divergence is indicative of a lack of agreement between the test-maker, content specialist or test user and the test-takers with regard to what topics the question is perceived to be about. The item-domain divergence measure may be viewed independently, and from the results of this study is mostly statistically independent from the responses to the questions that produced test scores for examinees. Discrepancies between $1 - (\phi/\phi_{max})^2$ across domains for the same item indicate that the inclusion of one of the domain components from the standard set in examinee's classifications is more highly correlated with responding correctly to that item than another domain in the "standard" classification. With respect to the set of simple descriptive statistical measures of the total test scores for those examinees who responded incorrectly and classified standard domains differently for the credit points of the test, the likelihood is greater that examinees with high total test scores misinterpreted the intended meaning of the items. Having higher total test scores, these cell C examinees were probably more likely to have conceptualized the intended meaning of an item like the teacher or content specialist than examinees with lower total test scores. Conversely, teachers or content specialists may not be able to detect interpretive defects in items as experienced by examinees of lower ability with lower total test scores. For those evaluating achievement, it is difficult, if not impossible, to disentangle interpretive defects in items from the more inadequate understanding the lower ability examinees have about the answers to the questions. Since these defects are more likely to be defects if higher ability students perceive the item to be about an unintended topic, these descriptive

statistics of total test scores would signal interpretive defects in the wording or meaning of items. These statistics provide clues for the re-examination of the originally intended meaning for items with relatively high values of $X_{tot,C}$ and $z_C$ in the form of a focus. This focus could take the form of a question. "What about the wording or structure of this question and the topic it was judged to be generated from cause higher ability examinees to respond to it incorrectly and not include this domain in their classification?" Although, $X_{tot,C}$ and $z_C$ provide clues when they take on relatively high values, defects in items were probably likely to affect lower ability examinees as well.

For the index of item-domain divergence to be effective in providing clues for interpretive defects in items, it must take on a relatively high value. For this to occur, $d_{ik}^{(D)}$ must not be stripped of its usefulness by "over-fitting" the domain to the actual wording of the item (e.g. the topic heading or domain label is included in the wording of the item or is identical to an acceptable answer to the question). The focus for re-examining the item provided by $d_{ik}^{(D)}$ is also best put in the form of a question. "What about the wording or structure of this question and the topic it was judged to be generated from cause so many examinees to not include this domain in their classification?"

For $1 - (\phi/\phi_{max})^2$ to serve as an effective clue for detecting technical defects in items, there must be a relatively large discrepancy between the values of $1 - (\phi/\phi_{max})^2$ for the domains that were judged to have generated the item across items. The focus for re-examining the item provided by such discrepancies can again be best put in the form of a question. "What

about the wording or structure of this question and its associated standard domains made for a higher correlation between responses and including one domain in their classification than the correlation between responses and including another domain?"

The response and classification data for Item 6 resulted in some examinees with high total test scores who responded incorrectly and did not include series circuit in their classification. Those of the eight examinees who had the higher total test scores classified the item as having been generated only from the domain parallel circuit which seems to have been related to how the responses were graded. However, these examinees were likely to have known the intended correct answer, but likely did not think it mattered if their responses were worded negatively thereby implying an advantage of a parallel circuit. Also, even though $d_{ik}^{(D)} = 0.113$ for "parallel circuit," there was one examinee with a total test score of 48.00 who responded incorrectly and did not include parallel circuit in their classification.

For Item 14, an interpretive defect very similar to the defect in Item 6 was perceived to exist. For the "main advantage" of a fuse being "current limiting," the topic, abnormal circuit conditions, conveys that current limiting is an abnormal circuit condition which is what a fuse is ordinarily used for. But, if a fuse is current limiting, it can protect equipment. There were a number of high ability examinees who responded with "current limiting" and classified the item as having been generated from ground-fault protection of equipment and not abnormal circuit conditions. $d_{ik}^{(D)} = 0.394$ for this item which is not too low to limit the

usefulness of the topic, abnormal circuit conditions, but is not so high as to become the primary signal for this item's ambiguity.

For Item 15, ambiguity seems to have revolved around the interpretation of the word "conditions." For one condition, there was one examinee in the incorrect-different category with a total test score of 49.00, and for the other condition, three examinees had a mean total test score of 46.00 with $s^2 = 3.00$ implying that the three scores clustered around 46.00. For this item $d_{ik}^{(D)} = 0.252$. However, this value was less useful or reliable than for Item 14 because the word "condition" appeared in the actual wording of the item and in the wording of the domain, abnormal circuit conditions. (This problem was previously encountered with Item 6).

Item 21 could be considered the most ambiguous since $d_{ik}^{(D)} = 1.068$ for circuit equations as compared to the high $d_{ik}^{(D)}$s for other items including "circuit equations" in their standard classifications. For this item, $d_{ik}^{(D)}$ was the highest for all components of items' standard classifications for the entire test. This indication of ambiguity is reinforced by the fact that there were enough examinees with high total test scores who responded incorrectly and classified differently to make the interpretive problem connected to the word "difference" more credible.

What about the responses to other items that produced high total test scores for those responding incorrectly and classifying differently? Item 2 has already been discussed in terms of a curricular validity problem Responses to item 4 produced one or two examinees with high total test scores who responded incorrectly and classified differently. For this

item, however, the components of the standard classification was less useful for detecting ambiguity. Item 12 has already been discussed in terms of curricular invalidity which may have accounted for examinees not making the connection between circuit equations and the solution to the problem. For Item 13, there were two cell C examinees with high total test scores. The value of this information, however, is offset by the fact that the topic, transformers, from which the item was generated is included in the wording of the item.

Just because items are deemed ambiguous from viable topics does not necessarily make them the most ambiguous. It may be that if more useful topics were used, other items may have been judged as ambiguous. However, without the information that could have been obtained, and since the interview data supported inferences about the possible misinterpretation of the wording, content or meaning of Items 6, 14, 15 and 21, only these inferences were made.

It may seem that the methods proposed are unwieldy and seem to be making "a mountain out of a molehill" with respect to the wording of items. But ambiguity is essentially a problem in communication. And if comparing the test maker, content specialist or test user's judgments about the topics an item is perceived to be about, then the potential communication gap is at least being addressed. In addition, the examination on which this research was conducted consisted of many different forms of questions. Many of them were open-ended in nature whereby a written response was graded. Multiple-choice tests, however, are more commonly used. And with multiple-choice tests, as they presently exist,

little can be known about how an item was interpreted without looking at examinee profiles across items via error-analysis of chosen distractors. For such tests, the reduction of response sets also reduces the potential for detecting item ambiguity without classification data. For this reasor, obtaining item-domain divergence measures and other statistical information through the use of item classification procedures can potentially be more beneficial to multiple-choice items. The benefit is the determination of the items that were most likely susceptible to misinterpretation assuming the measures are based on calculations for items whose classifications were non-trivial and useful.

Guidelines for the Improved Usage of the Proposed Methods for Detecting Item Ambiguity or Invalidity

Since individual components of some of the items' standard classification were deemed artifactually high, and because this innovative test technique as originally introduced by Smith and Tyler (1942) was not intended for statistical analyses, it was necessary to interview the examinees in order to determine potential issues concerning the items in the exam about which test takers and test makers differed. Each issue was discussed with respect to the classification data comprised of the item domain divergence measure, $d_{ik}^{(D)}$, and the item-knowledge type divergence measures, $d_{ik}^{(K1)}$ and $d_{ik}^{(K2)}$, and with respect to the classification information in combination with item response data for each item across examinees. Finally, judgments about the items were made from the plausibility of the reasons that are predicated to have produced the statistical

results. Inspection of item-domain divergence indices and the total test scores for those responding incorrectly and not including a standard domain in the classification across items is what determines evidence of item invalidity.

In order to guard against artifactually high divergence indices, it would have been better to instruct examinees that only one topic can be used for each item or make it particularly clear as to the number of topics used to generate each item. This would provide flexibility for item writers to construct items with as many domains in mind as are necessary and at the same time create less of a possibility for the occurrence of unwanted response sets for the behaviors associated with classifying items. The instructions for classifying should also make it clear that there is not necessarily an item that corresponds to every topic heading. From the problems encountered in making inferences in this research for assessing the validity of items in the interpretive sense, it would appear beneficial to not use any substantive content expression from the item or correct answer to the item in the actual wording of the topic heading or domain label unless absolutely necessary.

For reasons not quite determinable, the knowledge type classifications were not as useful as would have been desired for making inferences about the interpretive component of the validity of items. In addition, because of testing time constraints, the knowledge type classifications seemed impractical. It was conjectured that the knowledge type classifications were related to the wordiness of responses for open-ended questions in this examination. And very high values $d_{ik}^{(K1)}$ and $d_{ik}^{(K2)}$ occurred with

questions of a mathematical nature. It may also be that the teaching assistant's judgments were themselves not valid. If this is so, then the variability in the interpretation of the meanings for definition, fact, principle and interpretation that was provided for examinees produced these variable and mostly unsystematic indices of item-knowledge type divergence for the items on the test. In fact, this writer intuitively disagrees with some of the teaching assistant's knowledge type judgments. For these reasons, measures of item-knowledge type divergence should be employed for investigating other aspects of examinee item conceptualization phenomena rather than for detecting item ambiguity. It would be more practical if such categories were used during the examination without also having examinees classify the items according to topic.

Other Considerations and Viable Avenues for Future Research

A potential by-product of this procedure would be that it becomes more necessary for teachers to use a well-suited list of topic headings from a well-organized outline to generate items. If the instructions for test taking incorporate the classification of items with respect to such a well-suited list of topics, then students who would usually perform poorly on an exam, may learn more from the exam and perform more successfully. The topic headings by serving as advance or "concurrent" organizers could potentially improve test performance. In light of some of the examinees' comments, the possibility that classifying prior to answering may enhance examinee performance warrants further study for a variety of testing contexts and question types (e.g. multiple-choice, fill-in, oper-ended, etc.).

Future research on the use of item classifications may consist of determining whether or not classification of items enhances performance. Examinees from the same instructional grouping could be randomly assigned to either a testing situation that included instructions for examinees to classify or to a control group for taking the same test in the usual way. An offshot of this would be to randomly assign examinees to several treatments using the same test and where all groups would be instructed to classify the items in different ways. It would be hypothesized a priori that the group which was instructed to classify each item and then answer it would on the average receive higher total test scores. Another offshot of this research would entail instructing examinees that they would receive partial credit for their classifications in order to determine the extent to which examinees who usually perform poorly, know what the items are generally about. This would provide educators with more information about the state of knowledge each pupil has. Intervening variables like socio-economic status and ethnic background could be used as dummy variables in regression analyses to determine if there exists a cultural basis for knowing or not knowing what topic(s) questions are generally about.

What an item purports to measure is very similar to asking if examinees knew what the question was about enough to comprehend the question in the same general way as the test-maker, content specialist or test-user. This is tantamount to saying that on the average examinees know what the question is about. This study investigated the usefulness of a procedure whereby examinees classified test questions as to the topics or domains they were generated from in order to detect weaknesses or technical defects

in item wording or structure which may have led to examinee misinterpretation of item content. These misinterpretations represent examinee perceptions of items which are usually independent of subject-matter achievement or ability that is ordinarily inferred from test scores alone. Assessing the "interpretive" component of what an item purports to measure does not mitigate the importance of content and construct validity studies, but may aid in such studies as a "new variable of interest" for making more sound interpretations of test scores.

# REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C., 1974.

Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational Measurement (2nd Ed.), Washington, D.C.: American Council on Education, 1971.

Dawson, J.A. The validity of qualitative investigation. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Easley, J.A., Kendzior, E. & Wallace, R. A "bio-assay" of biology tests. The American Biology Teacher, 1967, 29, 382-389.

Glass, G.V., and Stanley, J.C. Statistical methods in education and psychology. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1970.

Hambleton, R.K. Validation of criterion-referenced test score interpretations and standard setting methods. Paper presented at the First Annual John Hopkins University National Symposium on Educational Research, Washington, D.C., 1978.

House, E.R. Technology and evaluation. Educational Technology, 1973, 13 (11), 20-26.

_____. Coherence and credibility. Educational Evaluation and Policy Analysis, 1979, 1, 5-17.

Lindquist, E.F., & Cook, W.W. Experimental procedures in test evaluation. Journal of Experimental Education, 1933, 1 (3), 163-185.

_____. Objective achievement test construction. Review of Educational Research, 1935, 4, 469-483.

_____. The theory of test construction. In H.E. Hawkes, E.F. Lindquist, & C.R. Mann (Eds.), The Construction and Use of Achievement Examinations. Boston: Houghton-Mifflin Company, 1936.

Linn, R.L. Issues of validity in measurement for competency-based programs. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1977.

Linn, R.L. Issues of validity in measurement for competency-based programs. In M.A. Bunda, & J.R. Sanders (Eds.), Practices and Problems in Competency-Based Measurement. National Council on Measurement in Education, 1979.

MacKay, R. Standardized tests: objective/objectified measures of "competence." In A.V. Cicourel (Ed.) Language Use and School Performance. New York: Academic Press Inc., 1974.

Messick, S.A. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Pines, A.L., Novak, J.D., Posner, G.J., & VanKirk, J. The clinical interview: A method for evaluating cognitive structure. (Research Report No. 6, Curriculum Series) Department of Education, Cornell University, Ithaca, New York, September 1978.

Polin, L., & Baker, E.L. Qualitative analysis of test item attributes for domain-referenced content validity judgements. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Popham, W.J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

Proppe', O. Personal communication, 1979.

Rovinelli, R.J. and Hambleton, R.K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.

Scandura, J.M., & Wells, J.N. Advance organizers in learning abstract mathematics. American Educational Research Journal, 1967, 4, 295-301.

Smith, E.R., Tyler, R.W., & the Evaluation Staff. Appraising and recording student progress. New York: Harper & Brothers, The Progressive Education Association Publications, 1942.

Stake, R.E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

# APPENDIX A

NAME (Professor's Key)

ET 242
Spring, '79
Closed Book (54 pts.)

On the line provided to the right of each item, place the letter(s) of the topic(s) that you think generated that item. The list of topics is projected on the overhead screen.

1. Below are examples of six electrical source types. Name each type. (6 pts.) A

   Photo-cell _____
   Battery _____
   Microphone _____
   Thermocouple _____
   Static Electricity _____
   Generator

2. T F A carbon-zinc battery is preferable to an alkaline battery when long life is important. (1 pt.) A,B

3. In order for currect to be present, two conditions must be met. What are they? (2 pts.) C

4. Name the three characteristics of a series circuit (answer in terms of currect, voltage, and resistance). (3 pts.) F

5. Name the three characteristics of a parallel circuit (in terms of I, V, R). (3 pts.) G

6. What then is the main advantage of a parallel circuit over a series circuit? (2 pts.) F,G

7. Schematically diagram a three-way switch. N

8.

A. Series B. Parallel

Wire the above diagrams so they form the appropriate circuit type. (2 pts.) F,G

9. What is the common method of reducing line losses between a power station and a building? (2 pts.) H.A

10. A 1200W hair dryer with a resistance of 12Ω requires a voltage of _____. (2 pts.) D,E

11. (6 pts.) Given a 12V system and the following different connections, E,F,G

    A. 6 - 20V Lamps in series
    B. 6 - 20V Lamps in parallel
    C. 9 - 20V Lamps in series

    A single lamp in _____ will burn the brightest.
    A single lamp in _____ will burn the next brightest.
    A single lamp in _____ will burn the least brightest.

    A single lamp in _____ has the longest life.
    A single lamp in _____ has the next longest life.
    A single lamp in _____ has the shortest life.

12. A. 6 - 20V lamps in parallel.
    B. 9 - 20V lamps in parallel. E,F,G

    Assuming line loss is negligible, compare life and brightness of a single lamp in the two systems. (2 pts.)

13. A step-up transformer with a 1 tp 30 winding ratio produces a secondary voltage of 100V. What is the primary voltage? _____V. (2 pts. H

14. Name the main advantage of the following (4 pts.): J

A. Fuse

B. Circuit Breaker

15. Name the two conditions which cause a fuse to blow or a circuit to trip and briefly explain each. (4 pts.) J

A. _____

B. _____

16. (4 pts.) What will happen to a circuit breaker if current: J

A. exceeds interrupting capacity?

B. exceeds full load current capacity?

17. Why does a ground fault circuit interrupter (GFCI) provide better protection than a circuit in which grounding is present? (2 pts.) L

18. _____ is measured as a difference in charge between two points. (1 pt.) B

19. _____ is a flow of charges. (1 pt.) C

20. A(n) _____ converts electrical power at one voltage to electrical power at a different voltage. (1 pt.) H

21. What is the difference between an electro-motive force (EMF) and a voltage drop? (2 pts.) B,E

# APPENDIX B

TOPICS COVERED BY TEST

A. Electrical Power System

B. Electromotive Force

C. Electric Current

D. Resistance

E. Circuit Equations

F. Series Circuit

G. Parallel Circuit

H. Transformers

I. Transformer Connections

J. Abnormal Circuit Conditions

K. Ground-Fault Protection of Equipment

L. Ground-Fault Protection of People

M. Lightning Protection

N. Graphics